CUG- H05949-24-1P 2687

THE JOURNAL OF PHILOSOPHY

VOLUME CIV, NUMBER 7
JULY 2007

page

321 *What Is Moral Judgment?* Richmond Campbell

350 *Self-Bias, Time-Bias, and the Metaphysics of Self and Time* Caspar Hare

374 NOTES AND NEWS

***Published by* The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Wolfgang Mann, Mary Mothersill, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C Varzi *Consulting Editors* David Albert, James T Higginbotham, Robert May, Charles D Parsons, Wilfried Sieg *Managing Editor* John Smylie

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President*, Arthur C Danto, *Vice President*, Akeel Bilgrami, *Secretary*, Daniel Shapiro, *Treasurer*, Barbara Gimbel, *Other Trustees* Lee Bollinger, Leigh S Cauman, Kent Greenawalt, Michael J Mooney, Lynn Nesbit

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

You may also visit our website at www journalofphilosophy org

THE JOURNAL OF PHILOSOPHY

2007

SUBSCRIPTIONS (12 issues)

Individuals	$45 00
Libraries and Institutions	$100 00
Libraries and Institutions (print and online)	$180 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

KP 2687

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for advertising rates, ad space is limited, so ad reservations are required

Published monthly as of January 1977, typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027 (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U S copyright law All rights reserved For more information about our copyright policies, please consult our January issue or contact us

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc , MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027 Periodicals postage paid at New York, NY, and other mailing offices

POSTMASTER. Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027

© Copyright 2007 by the Journal of Philosophy, Inc ISSN 0022–362X

THE JOURNAL OF PHILOSOPHY
VOLUME CIV, NO 7, JULY 2007

WHAT IS MORAL JUDGMENT?*

We speak, reason, and feel as if moral knowledge is possible, yet this idea confronts a serious, well-known problem The problem is so central to meta-ethics that one philosopher regards it as *the* moral problem in ethical theory [1] Moral knowledge appears to require moral judgments to be states of belief, yet they must at the same time be states of desire and feeling if they embody the motivation that we feel when we make moral judgments How can the same judgment be a state of belief and a state of desire or feeling, simultaneously?

More exactly, the problem is the logical incompatibility between three intuitively plausible propositions about (the nonempty[2] domain of) moral judgments [3] First, there is the seemingly unobjectionable view

(1) Moral judgments are (or express[4]) states of belief [5]

*I thank Darren Abramson, Pete Bahr, David Braybrooke, Wayne Fenske, Rockney Jacobsen, Andrew Kernohan, Victor Kumar, Duncan MacIntosh, William Rottschaefer, Greg Scherkoske, Meredith Schwartz, and Susan Sherwin for constructive criticism

[1] Michael Smith, *The Moral Problem* (Cambridge Blackwell, 1994)

[2] On the usual interpretation of universal propositions, (1), (2), and (3) are inconsistent only if there are moral judgments That we make moral judgments, given a neutral characterization of them, has never been in dispute and I will assume it throughout

[3] I intend 'judgment' to be neutral between cognitivist and noncognitivist interpretations of moral thought Readers who prefer 'thought' or 'response' or some other term for this purpose can replace 'judgment' throughout Nothing depends on its precise connotations

[4] 'Judgment' can refer to either a mental state or language expressing that state I focus on the first case and discuss judgments as linguistic expressions only in section VI I argue there that the problem and its resolution are essentially the same either way

[5] Beliefs always have truth value as I use the term, but noncognitivists are free to deny that moral judgments are ever beliefs in this sense

0022-362X/07/0407/321–49 © 2007 The Journal of Philosophy, Inc

Anyone who holds that moral judgments sometimes embody moral knowledge would be committed to this proposition on pain of denying that moral knowledge entails moral belief Unfortunately, this commonsense view conflicts with two widely accepted views of moral motivation often attributed to Hume [6] First, anyone who genuinely makes a moral judgment about something feels positively or negatively about what is judged in virtue of the intrinsic normative force of the judgment and is motivated (to some degree) to be guided by the moral judgment, independently of any antecedent desires or feelings that may also be sources of feeling and motivation More simply

(2) Moral judgments always move the judger intrinsically, independently of antecedent desires and feelings

Second, mere beliefs, independently of desires and feelings, never motivate, when beliefs move us, they do so because we also have antecedent desires and feelings and our beliefs elicit our responses, given those desires and feelings Again, more simply

(3) Beliefs never move the believer intrinsically, independently of antecedent desires and feelings

Obviously, if there are any moral judgments, these three views taken together lead to contradiction The problem for the defender of moral knowledge is thus to explain which of the two Humean tenets about the nature of motivation should be surrendered It is at the same time a problem about making sense of moral judgment, since we do make moral judgments yet our conception of moral judgment in relation to belief and motivation appears to entail a contradiction [7]

This problem may be resolved, I shall contend, by understanding moral judgments to be complex, multifunctional states that normally comprise *both* states of belief that represent possible moral truths *and* states of emotion and motivation [8] The position that I propose is not

[6] Hume, *A Treatise of Human Nature*, L A Selby-Bigge, ed (New York Oxford, 1978), pp 413–18, 455–76

[7] The problem can be stated also by replacing (2) and (3) with "Moral judgments are, just by themselves, reasons to act" and "Beliefs are never, just by themselves, reasons to act " Since 'reasons to act' must mean motivating reasons to act for the claims to be plausible, the difference is largely verbal and the resolution that I propose works for both versions

[8] I first floated this view in my *Illusions of Paradox* (Lanham, MD Rowman and Littlefield, 1998), pp 169–75 I apply a preliminary version of it to Hume's moral philosophy in "The Problem of Moral Judgment," in Susan Sherwin and Peter K. Schotch, eds , *Engaged Philosophy Essays in Honour of David Braybrooke* (Toronto University Press, 2007), pp 249–69

to be confused with the problematic view that moral judgments are ontologically unique states ("desires"[9]) that are *necessarily* both beliefs and desires I contend, on the contrary, that a naturalized, multifunctional understanding of moral judgment offers a defensible alternative to "internalist" accounts of moral judgment (such as the desire theory) as well as to the standard "externalist" treatments Moreover, it offers an alternative perspective on the cognitive role of moral emotion in moral knowledge In sum, I oppose the false opposition between reason and emotion that has been with us since Plato and has not been resolved by either internalist or externalist versions of moral realism

I will be taking moral realism as the default position in ethical theory, since it agrees with common sense (we speak and think as if moral knowledge is possible) and is backed by powerful theoretical considerations that are amply rehearsed in the literature [10] However, the understanding of moral judgment that I defend is neutral between moral realism and error theory, the view that moral judgments have truth value but none are true due to a systematic error that is necessarily a part of moral thinking [11] The reason is that moral realism and error theory face the same conundrum They are both committed to cognitivism and hence must cope with the inference from plausible premises (2) and (3) to the denial of (1) [12] If my understand-

[9] The term is introduced in J E J Altham, "The Legacy of Emotivism," in Graham MacDonald and Crispin Wright, eds , *Fact, Science, and Morality* (New York Blackwell, 1987), pp 275–88, on p 284 Margaret Olivia Little subtly defends the desire version of internalism in her "Virtue as Knowledge Objections from the Philosophy of Mind," *Noûs*, XXXI (1997) 59–79

[10] David Copp elaborates these reasons in *Morality, Normativity, and Society* (New York Oxford, 1995), pp 15–19 A complementary understanding of moral realism is given by Geoffrey Sayre-McCord, "Introduction The Many Moral Realisms," in Sayre-McCord, ed , *Essays on Moral Realism* (Ithaca Cornell, 1988), pp 1–23 I review the literature in my "Moral Epistemology," in Edward N Zalta, ed , *The Stanford Encyclopedia of Philosophy* (Spring 2003 Edition on the Internet)

[11] J L Mackie, *Ethics Inventing Right and Wrong* (London Penguin, 1977), Richard T Garner, "On the Genuine Queerness of Moral Properties and Facts," *Australasian Journal of Philosophy*, LXVIII (1990) 137–46, Stephen Schiffer, "Meaning and Value," this JOURNAL, LXXXVII, 11 (November 1990) 602–14, Richard Joyce, *The Myth of Morality* (New York Cambridge, 2001), Hallvard Lillehammer, "Moral Error Theory," *Proceedings of the Aristotelian Society*, CIV (2004) 93–109, Mark Kalderon, *Moral Fictionalism* (New York Oxford, 2005)

[12] Kalderon's error theory may be an exception Although he takes the content of moral judgment to have a truth value and to be false, acceptance of a moral judgment is given a noncognitivist interpretation Nevertheless, if his theory implies that moral judgments are not beliefs, it goes against the intuitions backing (1), and if it implies (1) then, whether or not we call it "moral cognitivism," the theory evidently contradicts (2) and (3)

ing of moral judgment resolves the problem for moral realism, it does it for error theory as well [13]

I AGAINST EXTERNALIST MORAL REALISM

A well-established strategy for defending moral realism is to argue that (2), despite its evident appeal, is really false and that moral realism need not accept it The connection between moral judgment and felt motivation, it is argued, is mediated by the presence of antecedent desires and feelings that are "external" to the judgment itself In fact, the naturalistic accounts of moral authority characteristically adopt this externalist perspective on the nature of moral motivation and feeling [14] I will argue, however, that the externalist solution to the problem of the inconsistent triad is not satisfactory, at least not as this approach is usually conceived, and therefore a different solution must be found for moral realism

What is externalism? The distinction between "internalism" and "externalism" was introduced in 1947 by W D Falk to distinguish between two opposing conceptions of the relation between moral judgments and motivation [15] While internalism conceives their relation to be necessary, externalism denies that it is [16] But to think of the distinction as comprised of two mutually exclusive and exhaustive views of moral judgment leaves out the positive view of externalism contained in contemporary accounts Externalists reject the internalist claim that the relation between moral judgment and motivation is necessary because they conceive of moral motivation being mediated by desires and feelings that exist prior to the moral judgment and that explain instrumentally why moral

[13] It should be noted the problem for error theory is especially acute if error theory incorporates Mackie's reasoning that moral judgments imply that there exist properties that are intrinsically motivating and thus imply what is ontologically unacceptable from Mackie's perspective, for then his theory implies (2) and (3) At least it does if we include in his scientific perspective Bayesian theories of motivation and recent work on brain functioning as analyzed in Timothy Schroeder, *Three Faces of Desire* (New York Oxford, 2004), pp 157–61

[14] For example David Gauthier, *Morals by Agreement* (New York Oxford, 1986), Peter Railton, "Moral Realism," *The Philosophical Review,* XCV (1986) 163–207, Copp, *Morality, Normativity, and Society*

[15] Falk, "Ought and Motivation," *Proceedings of the Aristotelian Society,* N S XLVIII (1947–48) 11–38

[16] Henry Sidgwick, *The Method of Ethics* (Chicago University Press, 1907), William K. Frankena, "Obligation and Motivation in Recent Moral Philosophy," in A I Melden, ed , *Essays in Moral Philosophy* (Seattle Washington UP, 1958), pp 40–81, David O Brink, *Moral Realism and the Foundations of Ethics* (New York Cambridge, 1989), Sigrun Svavarsdottir, "Moral Cognitivism and Motivation," *The Philosophical Review,* CVIII (1999) 161–219

judgments normally have motivating force For example, if what is just promotes what we antecedently care about, the fact that something is just gives us an instrumental reason to care about it Conceding that moral judgments motivate, externalism understands the source of moral motivation to lie outside the moral facts themselves and thus outside the content of moral judgments The two views are thus contraries rather than contradictories It is possible, at least logically, for both to be false, since it is possible for the relation between judgment and motivation to be neither necessary nor merely instrumental

Externalists employ two strategies to establish their position One is to argue that the relation in question is not necessary and infer that externalism is true since it explains moral motivation without implying that the relation is necessary The other is to defend the instrumental explanation of moral motivation, noting that it excludes internalism As it turns out, neither strategy is successful in avoiding inconsistency

The first strategy relies on imagining cases where a moral judgment is true (or at least is sincerely believed to be true) but the judger has no motivation of the required kind Insofar as such cases are possible the relation between judgment and motivation cannot be necessary A standard example is that of the amoralist who concedes that what he has done is morally reprehensible but just does not give a damn [17] Thus, Thrasymachus urges those who can get away with it to be unjust, knowing full well exactly what it is to be just but having no desire to be just [18] While such examples may be counterexamples to (2), this proposition can be modified in a way that completely nullifies the force of the examples and leaves (1) as problematic as before Suppose that normally, but not always, anyone who genuinely makes a moral judgment feels its intrinsic normative force and is motivated (to some degree) to be guided by the moral judgment, independently of any other desires or feelings that the person may possess More simply

(2*) Normally, moral judgments move the judger intrinsically, independently of antecedent desires and feelings

Notice that (1), the presupposition of moral realism, is just as inconsistent with the (2*) and (3) as it is with (2) and (3), but the

[17] Brink, *Moral Realism*, pp 46–47

[18] Nicholas Sturgeon argues convincingly that it is difficult to make sense of such amoralism without supposing that there are moral facts, see "What Difference Does It Make Whether Moral Realism Is True?" *The Southern Journal of Philosophy*, XXIV Supplement (1986) 115–41 For reasons directly below, accepting this point does not resolve the problem of inconsistency for moral realism

amoralist presents no problem for (2*) Moreover, (2*) being logically weaker than (2), can be no less plausible than (2) In sum, the standard attack on the internalism does nothing to remove the threat of inconsistency facing moral realism The problem of apparent inconsistency evidently has its source in something other than the idea of a necessary connection between moral judgment and moral motivation Unless externalist moral realism can deal with this deeper problem, it does not offer a viable alternative to internalism and its putative refutation of internalism is to no avail

What about the second strategy for defending externalism? As we have already noted, externalism is more than the denial of internalism It involves a positive thesis that the relation between moral judgment and motivation is instrumental A representative statement of the thesis is provided by Brink in the course of defending moral realism

> Whether the recognition of moral facts provides reasons for action depends upon whether the agent has reason to do what morality requires But this, of course, depends upon what morality requires, i e upon what the moral facts are, and, at least on standard theories of reasons for action, whether recognition of these facts provides reason for action will depend upon contingent (even if deep) facts about the agents desires or interests [19]

The instrumental nature of the relation between recognition of moral facts and the agent's desires and interest is well illustrated by Peter Railton's discussion of facts and values

> To deny Hume's thesis of the practicality of moral judgment, and so remove the ground of his contrast between facts and values, is not to deny that morality has an action-guiding character Morality surely can remain prescriptive within an instrumental framework, and can recommend itself to us in much the same way that, say, epistemology does various significant and enduring—though perhaps not universal—human ends can be advanced if we apply certain evaluative criteria to our actions [20]

It is clear that the ends in question are antecedent to moral judgments Richard Boyd in his influential defense of externalist moral realism makes this point explicit

> Ordinary factual judgments often provide us with reasons for action, they serve as constraints on rational choice But they do so only because of our antecedent interests and desires If moral judgments are merely

[19] Brink, "Moral Realism and the Sceptical Arguments from Disagreement and Queerness," *Australasian Journal of Philosophy*, LXII (1984) 111–25, at p 114

[20] Railton, "Moral Realism," p 170

> factual judgments, as moral realism requires, then the relation of moral judgments to motivation and rationality must be the same [21]

Moral judgment, in sum, is taken by externalists to be a belief about things like justice that stand in an instrumental relation to other things for which the judger normally has antecedent desires and feelings Judging something to be just, therefore, normally motivates because the person making the judgment already cares about the consequences of just action

Notice that externalism, so understood, contradicts not just (2) but also (2*) For this reason externalism appears to offer a way out of the deeper problem of inconsistency It does so, however, only if the instrumental thesis is true I believe that it is not, since it conflicts with reasonable views about moral evolution and moral development

Set aside for the moment the issue of what a moral judgment is What is critical for externalism as conceived here is whether the feelings and motives that bring us to act in accord with moral judgment do so on a merely instrumental basis Many theories exist about how moral emotion and motives might have evolved among humans based on Darwinian natural selection [22] What I take to be at their core is the thesis that pre-humans had better chances for survival (and reproduction) having moral feelings and motives (or something like them) than not having them This minimalist claim goes no distance towards explaining the evolution of particular moralities or even particular types of moral-like dispositions to act All it explains at best is why moral emotion and motivation (or something like them) came to exist or persist [23]

[21] Boyd, "How to Be a Moral Realist," in Sayre-McCord, ed , *Essays on Moral Realism*, pp 181–228, at p 186

[22] Most evolutionary accounts attempt to explain the origins of morality, such as Jessica C Flack and Frans B M deWaal, "'Any Animal Whatever' Darwinian Building Blocks of Morality in Monkeys and Apes," in Leonard D Katz, ed , *Evolutionary Origins of Morality* (Bowling Green, OH Imprint Academic, 2000), pp 1–29 An alternative evolutionary approach that is particularly appropriate for cultural evolution is epidemiological, looking not at the origins of norms but at why some norms prevail rather than others, see Dan Sperber, *Explaining Culture* (Cambridge Blackwell, 1996), and Shaun Nichols, "On the Genealogy of Norms A Case for Emotion in Cultural Evolution," *Philosophy of Science*, LXIX (2002) 234–55 The minimal assumption about moral evolution set out in this paper allows either approach

[23] I defend this view in "Can Biology Make Morals Objective?," *Biology and Philosophy*, XI (1996) 21–31 See Alexander Rosenberg, "The Biological Justification of Ethics A Best-Case Scenario," *Social Philosophy and Policy*, VIII (1990) 86–101, for problems with less minimal attempts at explaining the justification of norms via evolution A proposal about how motivations to act, triggered by natural features of things, evolve by natural selection to become moral reasons to act is given in Richmond Campbell and Jennifer Woodrow, "Why Moore's Open Question Is Open The Evolution of Moral Supervenience," *The Journal of Value Inquiry*, XXXVII (2003) 353–72

But this minimal evolutionary thesis is all I need, given certain very weak assumptions, such as the capacity for developing moral-like motivation and feeling is gene-linked and variation in this capacity occurred when it was important for survival These assumptions are not beyond question, but they are plausible enough not to call for a special defense

Even so the core theory may not appear to bear directly on the instrumental nature of moral feeling and motivation implied by externalism After all, externalists do not deny that moral motivation and feeling exist or that they conduce to survival Notice, however, that a purely instrumental story of why we tend to act in accord with our moral judgment does not fit even the minimalist account of how moral emotion and motives evolved and is unlikely to explain fully what sustains morals today That is because the purely instrumental story makes the motivating power of moral judgment too indirect for survival to depend just on it Other things being equal, natural selection will favor the more reliable means to survival and reproduction, and an indirect mechanism would be to some degree less reliable than the sort described (at the level of phenotype) in (2) or even (2*) [24]

A purely instrumental mechanism would be less reliable, because its effectiveness would depend on an agent's having true beliefs not only about what to do but also about the causal consequences of doing it Therefore, the reliability of the moral emotion and motives in causing us to do what we judge right would increase if we were moved to do it just because we perceive it as being right and this motivation did not depend always on perceiving in addition the right causal connection to our antecedent desires and interests In matters of natural selection incremental differences in reliability can make all the difference, when other things are equal Other things include the energy costs of the more reliable device and the availability of it (given constraints of mutation and development) The energy costs of a mechanism that directs an agent toward one end should not be significantly different from the costs when another end is the target Moreover, once emotional and motivational mechanisms are on the scene anyway, the question of availability is not an issue I conclude that the externalist theory of

[24] Here I adapt the evolutionary argument made by Elliott Sober and David Sloan Wilson in favor of pluralism in the debate about whether we are intrinsically motivated solely by pleasure, see Sober and Wilson, *Unto Others* (Cambridge Harvard, 1998), chapter 10

how we are moved morally does not fit well with even very basic evolutionary considerations

Three caveats are in order First, normally we are moved intrinsically (to some degree) not only to do what we think is right but also to do things that have (nonmoral) features that we take, other things being equal, to make something right, such as, helping people in need, telling the truth, keeping one's word, and being loyal But, again, for our moral feelings and motives to be sufficiently effective in moving us to do these things, we need to be moved intrinsically to do these things, not merely to be moved do them because their perceived consequences satisfy our antecedent feelings and desires Second, it would be folly to deny that moral motivation is in part instrumental At least by the time we are mature, we normally care about many things that honesty and justice, when practiced generally, tend to promote indirectly, such as the general well-being of others What I am arguing against is the implausible view that the attraction to being moral and to being just and honest is solely instrumental Third, it does not follow immediately that moral motivation is today partly noninstrumental given that moral-like motivation was so for early hominoids Yet it would be entirely gratuitous to suppose that immoral impulses are now relatively weak and we have such strong instrumental reasons to be moral that the capacity for noninstrumental moral motivation is useless or that we can now survive without moral motivation Neither of the latter hypotheses is at all likely

Developmental considerations reinforce evolutionary ones and constitute in their own right an argument against the purely instrumental story We know that early in their development children learn to feel negatively about categories of things that parents teach them are wrong The psychological means of teaching are blatantly instrumental but the resulting motivation is not Parents punish providing a negative (instrumental) motive to be good and children naturally want to please their parents so that they also have a positive (instrumental) motive to be good [25] Moreover, they tend to imitate their parents, being good in ways that they are in order to be like them (again an instrumental motive) But the result over time is that children come to "internalize" the norms that their parents try to instill in them with the result that they normally desire (to some degree) to be good because they find being good attractive in itself They are, that is to say, motivated to be moral noninstrumentally,

[25] Lawrence Kohlberg, "The Development of Children's Orientation toward a Moral Order 1 Sequence in the Development of Moral Thought," *Vita Humana,* VI (1963) 11–33

despite the instrumental character of the learning process [26] If this result is not achieved, so that the child only does what is called for when she cannot get away with doing otherwise or only in order to please or imitate, then the learning process is incomplete and the child is in a significant way handicapped in coping with her social environment Without the development of a moral conscience she will be not be counted on to behave morally except in special circumstances and various things of value, such as being trusted to make her own choices in life, will be foreclosed These developmental facts are commonplace and entirely consistent with scientific theories of desire that see desire as connected only contingently with belief [27] Noninstrumental moral motivation, in short, is central to normal moral development, however contingent its relation to moral belief

That said, there is no question that a child does not forget the instrumental reasons to be moral, and as adults we are still moved to be moral in order to avoid punishment, to be well regarded, and to be like those we admire We normally develop an empathy with others and are therefore naturally inclined not to cause gratuitous pain As a pluralist I have no reason to deny any of this, but I believe that the evidence is very strong that people normally have also noninstrumental reasons to do what they believe is right and to be moved intrinsically, for example, not to cause gratuitous pain even when they lack empathy and feel the impulse to cause pain To be precise, contrary to externalism, (2*) is evidently true Anyone who makes a moral judgment normally feels its intrinsic normative force and thereby is motivated (to a degree) to be guided by the judgment, independently of any other desires or feelings

II AGAINST INTERNALIST MORAL REALISM

If (2*) is acceptable and the moral realist is committed to (1), then it would seem that the moral realist's only option is to reject (3), given the apparent inconsistency among the three propositions This is the option the internalist realist takes [28] She understands the relation of

[26] See Schroeder, *Three Faces of Desire*, pp 146–50, on how intrinsic desires are learned, including the learning of intrinsic desires based on instrumental desires

[27] Though Schroeder's discussion of the neurobiology of desire in *Three Faces of Desire* supports only a contingent connection between moral belief and the desire to do the right thing (pp 157–61), the subject is noninstrumental desire, including the moral case

[28] Defenses of different forms of realist internalism include John McDowell, "Are Moral Requirements Hypothetical Imperatives?" *Proceedings of the Aristotelian Society*, Supplement LII (1978) 13–29, Mark Platts, "Moral Reality," in *Essays on Moral Realism*, pp 282–300, David McNaughton, *Moral Vision* (New York Blackwell, 1988), James Dreier, "Internalism and Speaker Relativism," *Ethics*, CI (1990) 6–26, and Little, "Virtue as Knowledge Objections from Philosophy of Mind "

moral judgment to moral motivation to be not only noninstrumental but also necessary, with the consequence that moral beliefs are not motivationally inert, when taken just in themselves, contrary to (3) The burden of this section is to determine if such internalism, committed to the denial of (3), is credible Do moral beliefs alone necessarily result in moral motivation?

The question is complex because moral realists of this stripe can espouse any one of three kinds of internalism One claim is that states of moral belief and states of moral motivation are distinct mental processes (or events or states) but that the first entails the second This form of internalism is evident among moral realists who insist that moral judgments are merely states of belief David McNaughton, for example, holds this view, since he describes moral judgments as "purely cognitive" states of belief yet understands the judgment to entail motivation on the part of the agent

> To be aware of a moral requirement is, according to the realist, to have a conception of the situation as demanding a moral response The requirement will only be satisfied if the agent changes the world to fit it But the realist also wishes to insist that the agent's conception of the situation is purely cognitive That is, the agent has a *belief* that he is morally required to act and so his state must have the direction of fit this state must fit the world [29]

The first direction of fit is that appropriate to desire, the second to belief, according to McNaughton, and a moral judgment entails both, though it is itself purely a belief

A second kind of internalism maintains the relation of entailment but holds that the moral belief and moral motivation are not separate mental states or processes but are both part of moral judgment A moral judgment on this view is essentially both a belief and a desire, or to use the term coined by J E J Altham, it is a "besire" (*op cit*) We have already noted an objection that would apply to both versions of internalism, namely that the amoralist is not beyond possibility Michael Stocker gives a number of everyday examples [30] A common rebuttal, that in these cases no genuine moral judgment is made,[31] has not been convincing, even for all committed internalists James Dreier, for example, would maintain that while these may be cases of

[29] McNaughton, *Moral Vision*, p 109, emphasis in the original

[30] Michael Stocker, "Desiring the Bad An Essay in Moral Psychology," this JOURNAL, LXXVI, 12 (December 1979) 738–53

[31] R M Hare initiated this line of objection in *The Language of Morals* (New York Oxford, 1952), pp 124–26, 163–65

moral judgment, the agents fail to be motivated because their conception of what they think is "morally required" is cognitively deficient The agents in question are abnormal, not merely statistically but also normatively, in that they lack the proper cognitive moral development needed for moral belief to issue in moral desire Thus, a third form of internalism "posits a necessary connection between believed good and motivation only in normal cases "[32] Given that the examples of amoralists used to defeat internalism are abnormal (in both senses of the term), these putative counterexamples are disarmed and the thesis that moral belief necessitates desire in cases where the agent has sufficient cognitive wherewithal remains in tact

Is there an objection that would defeat all three kinds of internalism? It is sometimes maintained that the necessary connection cannot hold in any cases because beliefs and desires have different "directions of fit" beliefs should fit the world, the world should fit desires [33] How can anything have both directions of fit at once? But this question is easily answered, since in the case of moral judgment the object of belief and the object of desire are distinct A moral belief, for example that slavery is wrong, is about slavery being wrong, while the moral motivation that goes with the belief is about stopping slavery [34] Moreover, it is not implausible to suppose that beliefs and desires can have the same object Ruth Millikan maintains that many states of mind are like this [35] Intending to go to the store involves both the motivation to go and the belief that you will go If there is a general problem with realist internalism, we need to look elsewhere

Transition cases, as I will call them, present a general problem that infects all three variations of internalism In some of these cases a person changes moral belief while still in the grip of feelings and desires learned when developing the earlier belief Imagine a person whose parents have raised him to regard gay sex as immoral because it

[32] "Internalism and Speaker Relativism," p 11 Note that on Dreier's interpretation, (2[r]) would be a conceptual truth It would imply that if one is normal in his normative sense (has no cognitive defect in moral perception) and makes a moral judgment, the judgment is necessarily intrinsically motivating I argue below that (2*) is false when interpreted in this internalist way, but it does not follow that (2*) is false when interpretated as merely a non-conceptual, statistical truth So interpreted, (2*) appears to be true and compatible with the theory of moral judgment that I advance to resolve the apparent inconsistency among (1), (2*), and (3)

[33] The distinction is due to G E M Anscombe, *Intention*, 2nd edition (Ithaca Cornell, 1963), p 56

[34] Little, "Virtue as Knowledge," pp 63–64

[35] Ruth Garrett Millikan, "Pushmi-pullyu Representations," in Larry May, Marilyn Friedman, and Andy Clark, eds , *Mind and Morals* (Cambridge MIT, 1996), pp 145–61

is unnatural and also contrary to the teaching of the Bible His belief and accompanying feelings of aversion and discomfort as well as the desire to eradicate this evil practice have been strongly reinforced by peers and teachers in his formative years However, after he has traveled away from home and met people with very different moral and religious views, he begins to question his religious faith and the attitudes that go with it He learns in biology class about the great variety of sexual response in nature and, after some serious reflection, forms for the first time the opinion that gay sex is not wrong It would not be unusual if the feelings and desires appropriate to his former belief remain with him for a period of time, however brief He finds himself believing that gay sex is okay but at the same time feels uncomfortable and embarrassed by it and feels a strong impulse to show his displeasure about it despite his new moral outlook On discovering that his child is gay, his immediate response may be negative and have no flicker of a desire to accept the situation or act in accord with his belief Here is a person in transition in his moral judgment in a situation where his feelings and desires lag behind his moral understanding, at least for a period of time

A different kind of transition case would be a woman who has been treated unfairly but fails to perceive any injustice given how she has been encouraged to think about this type of situation Perhaps she has been passed over for a job despite her being more qualified than the man to whom it is offered She does, however, feel intensely angry when she thinks about what has happened The problem is that she fails to see how she has been wronged in any way She believes that it is appropriate that men should be offered work before equally qualified women, because of the different roles that men and women have in society As in the previous example, we may imagine that her belief accords with her religious upbringing and both her belief and her earlier tendency to feel and want to act accordingly have been strongly reinforced However, as her experience expands and she comes to have more invested in making decisions for herself and acting as an independent person, her moral belief about this situation can come to conflict with her emotional and motivational response She continues to believe for a time that there is really nothing wrong even though she finds herself feeling and wanting to act as if she has been treated unjustly

In the first kind of transition case moral belief changes without an immediate change in feeling and desire, in the latter kind, feeling and desire change without an immediate change in belief Let us suppose further (though there is no necessity in this) that the moral response (feeling/desire or belief) that does not change initially

changes eventually (for reasons taken up in section v) and the transition from the old moral perspective to the new is complete Moreover, let us grant to the moral realist for the sake of argument that in reality gay sex is not wrong as such and that in fact the woman was treated unjustly Given these assumptions the cases can be described as someone's feelings, desires, and beliefs responding in an appropriate manner to moral reality, though responding not all at once but in a step-wise fashion I claim that transition cases, so conceived, constitute damaging counterexamples to all three forms of realist internalism

The force of the objection turns on two features of transition cases that are worth highlighting The first is that while these cases are abnormal statistically (most people's moral responses are fairly stable), the cases are not abnormal in the normative sense that would disqualify them from being counterexamples to internalism The second feature is that in the transition cases an aspect of the agent's moral judgment goes against what the agent has been taught to recognize as wrong These features make the standard defenses against counterexamples irrelevant The irrelevance in the first type of transition case is easy to see The agent appears to form a moral belief without immediately having the appropriate emotional and motivational state Since the belief goes against common thinking (we are supposing), the apparent moral belief cannot be discounted as meaning merely that this is what most people think Since the case is one in which the agent shows sufficient cognitive appreciation of moral reality, despite his upbringing, the realist internalist cannot discount the case as being cognitively subnormal But since the necessary connection between belief and motivation fails to hold for this case in which the agent sufficiently appreciates the nature of moral reality to form the right belief, the example runs counter to all three kinds of realist internalism

The irrelevance of the standard defenses in the second type of transition case should also be clear, though the argument is slightly subtler In this case the moral belief and the motivation of the agent are initially at odds, as before Again the necessary connection appears to fail But is this a case where the agent is cognitively deficient in the relevant way? The answer has to be no, because we are supposing that it is a moral fact that the agent is being treated unjustly and she has sufficient cognitive capacity to form an appropriate response, namely anger, despite the fact that this response goes against (we suppose) the thinking of most of her society Moreover, because her response is counter to those others, she cannot be fairly described as simply mimicking the feelings prevalent in society Indeed, in this

case a person appears to be making a moral judgment in virtue of making an appropriate cognitive response to moral facts despite the fact that she has not formed the corresponding moral belief I will elaborate the implications of this aspect of this second transition case in sections III and V

III MAKING SENSE OF MORAL JUDGMENT

Neurobiological considerations also tell against realist internalism,[36] but since the transition cases are reason enough to find an alternative, it is time to take stock If these cases block the internalist rejection of (3) and the externalist reasons for rejecting (2*) are blocked for the reasons given in section I, how can the moral realist avoid inconsistency? We appear to be left once more with the inconsistent triad (1), (2*), and (3) I believe that the answer lies in the interpretation of (1) the proposition that moral judgments are beliefs In the literature (1) is taken to express "cognitivism" and the denial of (1) to express "noncognitivism," as if these views were contradictories, leaving no logical room to reject both But (1), as stated, is multiply ambiguous The claim that moral judgments are beliefs could mean that they are only states of beliefs (hence not states of emotion and motivation), or it could mean that moral judgments are at least states of beliefs (and perhaps other states as well, either at once or on different occasions) Judging from the debate in the literature, one might be led to interpret (1) the first way and noncognitivism as the view that moral judgments are only states of emotion or motivation, making the views contraries But when (1) is read that way, it does not follow that both views cannot be false In fact, if moral judgments normally were both states of belief and states of motivation, as I shall now urge, then both would be false That is, both (1) read narrowly (moral judgments are only states of belief) and (2), the view that moral judgments are only states of motivation, would be false

[36] Schroeder argues in *Three Faces of Desire*, pp 157–61, that moral motivation based on representations of right and wrong is no different in its underlying neurobiological structure than other kinds of deliberative motivation The content is different but not the process leading to action Moral representations move us because of their projections on orbitofrontal cortex in the same way that representations of our favorite sandwiches move us He considers the possibility that moral representations project movement possibilities "directly into the motor centers, and so cause action, independently of desires Yet if moral representation would not merely promote possible actions, but actually override the inhibitory action of the basal ganglia, this would put moral motivation on a par with Tourettic urges and would make moral action a form of behavior tic" (p 160)

Does that mean that moral realism must be rejected because it implies (1)? No, (1) can be read broadly so that it allows that moral judgments are normally both kinds of states at once So understood, (1) would not contradict the conjunction of (2) and (3) Neither would it contradict that of (2*) and (3), and the argument against moral realism (in either of the versions above) would be defeated To avoid confusion, let us be explicit about the possible dual nature of moral judgments and substitute for (1)

(1*) Normally, moral judgments are both states of belief and of motivation

It is clear that (1*), (2*), and (3) form a consistent triad If, therefore, moral realism requires (1*), but does not require (1) read narrowly, then the threat of inconsistency vanishes

At this point a second ambiguity emerges (1), interpreted narrowly (or broadly), could mean that moral judgments are *always* (at least) moral beliefs But (1) could also be read the way we normally read "Humans have two legs" and "Birds fly" to imply that mostly they do This further ambiguity is important, since the second transition case suggests that moral judgments need not always be or include beliefs In other words, moral judgments come in three varieties some include only moral beliefs (first transition), some only emotional and motivational moral responses (second transition), and the vast majority (the paradigm on the dual view) include both Note that (1*) is compatible with this multifunctional view of moral judgment It even allows moral knowledge in the second transition case, as I will argue in section v

Are moral judgments genuinely dual in this way? Does moral realism require that they be so? There are two immediate reasons for affirmative answers to both questions The most fundamental is that construing moral judgments in this way dissolves the threat of inconsistency contained in the original triad of propositions Recall that we are assuming moral realism as the default position If we accept the hybrid understanding of moral judgment, moral realism can maintain the tripartite position that moral judgments are belief states (though not exclusively or invariably), that moral judgments normally involve moral motivation that is not dependent on antecedent states of desire or feeling, and that the belief states that are contained in moral judgments are not motivating just in themselves, independently of antecedent states of desire and feeling In other words, understanding moral judgments as having dual functions allows one to keep the original intuitions that moral judgment can embody moral knowledge, that moral judgment normally

moves the judger intrinsically, and that belief alone cannot move anyone in this way That this conception of moral judgment resolves this fundamental problem is a powerful argument in its favor, especially since this conception avoids the problems facing the standard alternatives

In fact, the position also explains the evidence to which each alternative makes appeal Its explanatory power is thus a second argument in its favor An attraction of realist internalism is its implication that moral motivation and feeling normally issues (in part) directly from the moral judgment itself, independently of antecedent desires and feelings The present proposal can explain this attraction because the proposal implies that moral judgment normally contains within it feelings and desires that directly move the person judging The dual character of moral judgment allows it to be intrinsically motivating On the other hand, the attraction of realist externalism is that it implies that the connection between moral belief and moral motivation, even in standard cases, is purely contingent, in accordance with our best psychological theory The present proposal, as we have seen, has the same implication In short, the proposal is able to enlist in its support the very considerations that have made the alternatives internalism and externalism attractive because it entails those considerations Moral judgments because of their dual nature have all the different virtues that are attributed to them by the competing theories but lack the features that make these theories implausible

IV THREE ALTERNATIVE THEORIES

Is there an alternative theory that we have not considered? In particular, what about Smith's quasi-internalist view of moral judgment? Does this modified internalism have equal explanatory power? On his view, moral judgments are about whether one would be motivated to do certain things if one were rational (*op cit*) To judge that famine relief is morally required in present circumstances is to judge that a rational person would provide famine relief in these circumstances Thus, if one is rational and this judgment is true, then necessarily one will provide famine relief Here the necessary connection between moral belief and moral motivation is mediated by rationality The proposal is ingenious in that it provides intrinsic motivation in the case of a rational person without directly violating the Humean dictum that beliefs by themselves are motivationally inert

It is useful to compare this view with the dual theory of judgment, since it faces several problems that the latter theory does not share

First, Smith's theory does not explain how false moral judgments motivate directly, though they should function the same way as true moral judgments according to any plausible account of how moral norms become internalized If one has internalized a norm prescribing capital punishment the dual theory implies that normally one will be motivated intrinsically to promote the death penalty for certain crimes even if (contrary to one's belief) the norms involved cannot be justified and one's moral judgments about such matters are false Smith's theory, on other hand, explains this kind of motivation only for cases where the judgment is made by a rational person and is true A second problem is that Smith's account implies that if one is rational one will be motivated in virtue of one's rationality to act morally At present there is no generally accepted account of rationality that has this implication The leading contenders are generally accounts that construe rationality as efficiency in satisfying antecedent desires and therefore do not logically require a rational person to act morally Here the dual theory offers a more plausible explanation of moral motivation Third, Smith builds into the content of moral judgment a specific understanding of moral justification a state of rational reflective equilibrium in judgments But people can make the same moral judgment on the basis of a great variety of views about what justifies the norms that they invoke, including different religious beliefs about the moral authority of God Whatever the merits of Smith's understanding of moral justification, a plausible account of moral motivation must not be held hostage to an implausible view about the content of moral judgment The dual account of moral judgment, by contrast, leaves the nature of moral justification open

Two other theories may appear to provide a genuine hybrid conception of moral judgment Michael Ridge proposes a theory called "ecumenical expressivism" in which "moral utterances express both desires and beliefs "[37] This view is expressivist rather than cognitivist, since it "gives logical priority to desire" (*ibid*, p 309) It does so because it takes moral predicates to express "both a speaker's attitude in favor of action in general insofar as they have a certain property (whatever property guides the speaker's approval of actions quite generally) and a belief which makes anaphoric reference to that property" (*ibid*, p 313) For example, a speaker may have a utilitarian

[37] Ridge, "Ecumenical Expressivism Finessing Frege," *Ethics*, CXVI (2006) 302–36 at 302

attitude that favors action just as long as it maximizes happiness In judging an action that has this property to be morally appropriate, the speaker is expressing her approval of it and her belief that it has the property that elicits her approval Her moral judgment is thus contains elements of both belief and motivation Ridge takes this view to escape the false dichotomy between the thesis that moral utterances express beliefs but not desires and the thesis that they instead express desires but not beliefs (*ibid*, p 305)

But does it? Moral belief on this theory is not inherently normative That a speaker believes that an action has a property that in general elicits her approval of the action and her desire to perform it, such as the property of maximizing happiness, is not a belief that is inherently normative according to this theory The way that Ridge puts the point is that the truth conditions of moral utterances are not necessarily provided by the beliefs that they express, no more than the utterance, "This action is morally appropriate," is necessarily true if it is believed to maximize happiness Given Ridge's expressivism, whether the property of maximizing happiness makes an utterance true is a first-order moral issue However, on the hybrid view that I am defending, the moral belief that is normally part of a moral judgment is inherently normative Thus, in the first transition case the moral belief state obtains in the absence of the emotional and motivational state that normally goes with the moral belief, but the believer is, nevertheless, making a moral judgment Such a case is not possible in ecumenical expressivism Belief alone, on this theory, has no moral status In short, the fundamental dichotomy between cognitivism and expressivism is preserved, however ecumenical the intention behind the theory

Ridge notes that one could advocate a view, ecumenical cognitivism, in which moral belief rather than desire has priority and cites David Copp as having the most developed version of it Like Ridge, Copp focuses on moral utterances and what they express [38] In Copp's view (which he calls "realist-expressivism"), a moral utterance expresses a moral belief and also "certain characteristic conative attitudes or motivational states" in virtue of semantic conventions His theory is not internalist, however, because a person expressing a moral judgment need not have the attitudes in question or be motivated in the way indicated by those conventions Put in my terms,

[38] Copp, "Realist-Expressivism A Neglected Option for Moral Realism," *Social Philosophy and Policy*, XVIII (2001) 1–43

there is no problem with the first transition case The moral belief expressed in the moral utterances can stand alone as an inherently normative moral judgment, in contrast to what is implied in ecumenical expressivism But this view as well is not genuinely hybrid, nor does Copp intend it to be Moral judgments expressed in moral utterances are essentially moral beliefs The attitudes and motivation that are expressed, though they are part of the judgment expressed when a moral belief is also expressed, do not constitute an aspect of moral judgment that could in the absence of moral belief be itself a moral judgment as in the second transition case The priority in the understanding of moral judgment is for Copp the belief state, not the state of attitude and desire In this way, Copp's theory too reinforces the traditional dichotomy, but in the opposite direction Neither his theory nor Ridge's is a fully hybrid theory in which both elements have equal status

V THE COGNITIVE ROLE OF EMOTION IN MORAL CHANGE

In a fully hybrid theory both parts of moral judgment, the belief and the state of emotion and motivation, have equal status, to the extent that each can function by itself as a moral judgment in the absence of the other The latter aspect allows the existence of transition cases that distinguish quasi-hybrids from the robust hybrid theory in which these parts of moral judgment are normally together but can sometimes stand alone The idea of moral belief being able by itself to constitute a moral judgment is not unfamiliar, we find it in externalist cognitivism Nor is the idea of a state of emotion and motivation being able by itself to constitute a moral judgment, it is the essence of expressivism But the idea of the latter state by itself constituting a moral judgment in the context of moral realism is not common How is it possible to treat such a state as being by itself a moral judgment without lapsing into expressivism?

In this section, I want to address the question by considering the role of emotion in moral judgment and then link this role with the content of moral belief I hope to show how the relation between moral emotion and the content of moral belief can explain why the parts of moral judgment normally function together to form a unified whole Thus far I have lumped together the elements of desire and emotion and speak of the theory of moral judgment as a dual or hybrid theory as if there are just two basic aspects The situation is of course more complex than this picture would suggest In particular, the role of moral emotion in moral judgment is different from that of desire Emotion, while normally a part of moral motivation, has a representational role in motivation that serves

to make beliefs more accurate and consequently allows emotion and belief, as guides to motivation, to function in a unified way in moral thought [39]

That emotion has an important part to play in moral thought is well recognized It functions, for example, to communicate moral attitudes through facial expression and body language more quickly and at times more accurately than verbal expression [40] Since morals depend for their effectiveness in coordinating behavior on people knowing how each other responds to moral issues, moral emotion has an important supplementary role to play in moral communication [41] Emotion also stabilizes moral judgments because it tends to be spontaneous and to a marked degree insensitive in the short term to changes in information [42] As in the second transition case, resentment can be unplanned and arise even in a person who believes that she has not been wronged Because of this feature moral attitudes rooted in emotion are not easily manipulated and are able to remain stable even when they run contrary to immediate self-interest At third important feature of moral emotion is its tendency to elicit similar feelings in others, even those who resist it, when the feelings are known to be already widely shared This feature complements the previous, though it may appear to work against it when considered in abstraction from social context Moral attitudes need to be stable but not too stable, or else social moral change is not possible [43] Again we have a feature of moral emotion that helps to promote coordinated responses to moral norms and stability in the process of moral change As important as these features are, however, they fit equally well the views of internalist and externalist moral realism, since neither denies that moral judgment involves in some way moral motivation and both can allow emotion to play a role in moral motiva-

[39] I am indebted to Meredith Schwartz for suggesting that the representational character of emotion allows moral emotion to play a unifying role in moral judgment, given my dual conception of it She develops her view in "Feeling Good The Role of Emotion in Naturalized Epistemology," presented on July 22, 2005, at Dalhousie University

[40] On the ability of facial expression to convey emotion, see Paul Ekman, *Unmasking the Face A Guide to Recognizing Emotions from Facial Expressions* (Englewood Cliffs, NJ Prentice Hall, 1975)

[41] Gibbard explores how emotion coordinates moral behavior in *Wise Choices, Apt Feelings A Theory of Normative Judgment* (Cambridge Harvard, 1990)

[42] See Paul E Griffiths, *What Emotions Really Are* (Chicago University Press, 1997), pp 91–98 An application to the problem of stabilizing moral attitudes is given in R H Frank, *Passions within Reason The Strategic Role of the Emotions* (New York Norton, 1988)

[43] Campbell and Woodrow, "Why Moore's Open Question Is Open," p 357

KP 2687

tion Indeed, the functions noted are also central to the story that noncognitivists tell about morals

How then does emotion make a difference? The difference resides in the fact that emotion has a representational function like that of belief [44] Consider fear, an emotion that is not specifically moral It is commonplace that belief and fear can share the same propositional object For example, I can both believe and fear that there is danger present It is also commonplace that both may fail to represent correctly However, in the latter case the belief, not the fear, is literally false Why is this? I suspect that we do not readily speak of fear as true or false because the primary function of motivational states is not to represent things as they are Whatever the reason, the similarity of fear to belief with respect to representation is striking Besides representing as fearful the same thing, fear like belief can be evaluated as being baseless or irrational depending on whether what is feared is true and whether the truth of what is feared can account for the fear If I fear that a spider will harm me but it cannot and I know it, my fear is unwarranted In another case I may feel fear because there is danger present and I have unconsciously responded to cues that give me good reason to be afraid, but for ideological reasons I do not believe that there is danger present I think of myself as being irrational and having groundless fear when in fact my fear is well grounded and, given the cues, rational Although the fear is not literally true, it can be evaluated in ways that are analogous to the way we assess the corresponding belief In particular, we can evaluate fear as being warranted or rational depending on whether what is feared is true and whether there is sufficient evidence for what is feared

Fear is not an exception in this respect Consider moral emotions like those that we discussed in the transition cases In one of these cases a person feels resentment for being passed over for promotion even though she believes that she has been treated fairly and not in any way wronged Is her emotion irrational and baseless? It may appear so given her belief that the process was legitimate and fair Yet it remains possible that she is mistaken about the latter and that her emotion is occasioned by events that indicate that she is mistaken

[44] In my *Illusions of Paradox*, pp 70–74, 173, I advance this view of emotion to support a form of moral realism in which moral knowledge can take the form of well-grounded moral emotion That view is compatible with the position taken here, but my focus now is on the relation of moral emotion to moral belief A like-minded conception of emotions is presented in Cheshire Calhoun, "Cognitive Emotions?," in Calhoun and Robert C Solomon, eds , *What Is an Emotion?* (New York Oxford, 1984), pp 327–42

(although she has ignored them or not given them appropriate weight) If the feeling of resentment persists, she may reflect on why she would feel this way and discuss her feeling with persons she trusts The outcome could go either way she may come to understand why she feels resentment and find that these reasons do not justify this reaction, or she may come to believe that despite her first impression there are good grounds for her feeling

The first point to notice is that the logical tension between the belief and the emotion is possible because the emotion has the power to represent something that (from a realist perspective) can be true or false The second point is that emotion, because of its representational power, is able to play a vital role in moral judgments changing to reflect moral reality more accurately Take the case of feeling disapproval or even revulsion regarding sexual behavior now believed to be entirely acceptable The emotion and the belief are in conflict because the moral disgust has the power to represent gay sex as morally repugnant If the belief that it is not has a more rational basis, the unified moral judgment that results can more accurately reflect moral reality than the disgust In the second transition case, because the emotion of resentment has a more rational basis than the belief that the treatment has been fair, the change is again to a unified moral judgment that more accurately reflects moral reality The representational character of the emotion allows a conflict between moral belief and moral emotion to initiate moral change and either to make the ensuing unified moral judgment more accurate and more rational

What are described in both kinds of change are types of moral transition that are not unusual and that afford a central role for emotion in moral change Another argument for the hybrid conception of moral judgment, then, is that it best explains role of emotion in the transition cases Moral motivation, including moral emotion, has equal status with moral belief on the dual conception Both are equally important in understanding rational change in moral judgment This equality in status is exactly what is manifest in the examples of moral transition given the role of emotion just described The second transition case, however, is especially significant Where the emotion is rational (that is, well founded) it can lead the way toward a more rational moral belief Externalist realism has severe problems already noted, but in addition, because it gives less than equal status to moral motivation, it cannot explain how emotion can play the role of transforming moral belief On the other hand, internalist realism cannot explain this role either, since it allows a split between motivation and belief only where the agent has

inappropriate motivation The case of rationally based resentment is just the opposite

Another way that the dual conception best explains the role of emotion in these cases is that it implies that moral judgment includes elements that form dynamic unity despite being functionally distinct By dynamic unity I mean that, while the elements can pull apart, as in the transition cases, and thus at times lack unity, the structure of moral judgment is such that elements will normally work together Given the role of emotion just described for these cases, we see that there is a natural tendency for the disunity to be resolved in favor of either the motivation or the belief Realist internalism may appear at first glance to offer something comparable, since it holds that moral belief explains the motivation that it normally necessitates But the connection is far from dynamic, since the moral motivation is supposed to be entirely determined by the moral belief with no room for the influence to go the other way or for moral motivation to be rational when it is opposed to moral belief

To summarize, transition cases are about change in moral judgment In contrast with the dual conception of moral judgment, both internalism and externalism overintellectualize moral change, though in different ways For externalism change in moral judgment is occasioned first by change in moral belief Because of her antecedent desires and feelings, an agent's moral motivation changes for instrumental reasons to accord with her change in moral belief[45] Moral motivation follows moral belief around like a puppy following its master, mostly sticking close but sometimes getting lost For internalism the picture is different but only superficially How one is motivated morally depends on what one believes, but the connection is necessary for agents who have sufficient cognitive capability The puppy never gets lost with a wise master In neither theory does moral emotion in conjunction with moral motivation transform moral belief, contrary to much everyday experience of change in moral judgment

[45] An exception is Copp's externalist realism in "Realist-Expressionism " In this theory an agent's motivation need not be purely instrumental because her motivation is integral to the expression of her moral belief (without being entailed by it) Still, in this theory emotion has no representational role to play in the creation of moral knowledge The theory cannot, therefore, properly explain moral change in which motivation and moral belief diverge because the emotion contained in moral motivation is in better touch with moral reality than moral belief For this reason it cannot explain the failure of those forms of realist internalism that do not require moral belief to entail moral motivation

So far I have said nothing about the content of moral belief, other than that it is able to express moral truth, and the arguments above do not depend on any specific view about moral content However, to better appreciate the unity among the different parts of moral judgment, consider what I take to be a reasonable hypothesis, namely, that moral beliefs imply that the states of emotion and motivation that would normally (but for the reasons given, not always) accompany them are justified in virtue of perceived features of the thing judged In other words, in the normal cases, what is believed when someone makes a moral judgment is that the same general type of emotion and motivation that would normally form a part of the judgment as a whole is in fact justified by perceived features of what is judged Consider Gilbert Harman's example of hoodlums setting a cat on fire for fun [46] If one suddenly comes on this scene and perceives the kids causing the cat extreme pain just for fun, one is apt to experience revulsion and anger, perhaps tinged with fear, together with an intense desire to stop the torture One also is apt to have the conviction, though it may not be consciously formulated, that this type of reaction just described is entirely justified by the fact that one is seeing a cat tortured for fun I want to suggest that this combination of ingredients constitutes a central case of moral judgment in which the belief involved is linked in content to the felt emotional and motivational response A full account would need to include an explanation of what makes someone's emotion and motivation generically moral, and more importantly, when features of what is judged justify a certain kind of emotional and motivational response I will not attempt to develop such an account here,[47] and the hybrid theory I defend does not depend on it Still, such an account would provide a yet deeper unity among the parts of moral judgment

VI A SEMANTIC OBJECTION

It may be objected that I have argued only for a unified theory of moral judgment understood as a psychological state and the arguments thus far do nothing to demonstrate that moral judgment understood in linguistic and semantic terms is similarly unified It may be objected further that the standard understanding of moral realism is semantic and that, unless the dual view of moral judgment can be extended to moral judgment taken as a moral claim expressed

[46] Harman, "Ethics and Observation," in Sayre-McCord, ed , *Essays on Moral Realism*, pp 119–24, at p 120

[47] An account of this kind is sketched in Campbell and Woodrow, "Why Moore's Open Question Is Open "

in language, no solution has been given to the inconsistent triad set out in semantic terms Finally, it may be noted that the apparently contradictory positions of cognitivism and noncognitivism are nowadays usually formulated just semantically [48]

To remove ambiguity let us use the term *moral judgment-E* to refer to the linguistic expression of a moral judgment (when it is used precisely to express a moral judgment)[49] and *moral judgment-M* to refer to the mental state of forming a moral judgment Then the corresponding inconsistent triad framed in semantic terms would be

(1#) Moral judgments-*E* express states of belief

(2#) Moral judgments-*E* always (or normally) express states of mind that move the judger intrinsically, independent of antecedent desires and feelings

(3#) States of belief never move the judger intrinsically, independent of antecedent desires and feelings

As before, I will assume that states of belief and their expressions are either true or false in order for (1#) to represent moral cognitivism Since moral realism, put in semantic terms, entails (1#) and thus is inconsistent with (2#) and (3#) and since the latter are plausible for much the same reasons as apply to (1) and (2), the new inconsistent triad constitutes an argument against moral realism

How different is this triad? It is easy to show that the old inconsistency among (1), (2), and (3) comes to the same thing if a moral judgment-*M* is nothing more or less than what is (or would be expressed) expressed by a moral judgment-*E* (I propose, as noted, to understand these terms so that what counts as a moral judgment-*E* depends in part on context) Assume that (1) is true, that all moral-judgments-*M* are beliefs Then every moral judgment-*E* expresses a state of belief Going the other way, assume that (1#) is true, that every moral judgment-*E* expresses a state of belief It follows that every moral judgment-*M* is a belief Assuming this equivalence between (1) and (1#), we see that (1#) contradicts the conjunction of (2) and (3) if the latter contradicts (1) and that (1) contradicts the conjunction of (2#) and (3#) if the latter contradicts (1#) In short, the difference

[48] Fifty years ago, theorists took noncognitivism to be a theory of morals that is both semantic and epistemic See Richard B Brandt's definition in *Ethical Theory* (Englewood Cliffs, NJ Prentice-Hall, 1959), p 205 Sayre-McCord provides a typical contemporary semantic formulation of cognitivism and noncognitivism in his "The Many Moral Realisms "

[49] 'Moral judgments-*E*', as I use the term, refers to moral assertions or claims, though the same words in another context might express doubt or disbelief

in idioms is superficial, which may explain why discussions of moral realism move freely between the psychological and the semantic

A true resolution of the problem, therefore, should fit both versions of the triad The resolution applied to the second case is to note the ambiguity in (1#) On a narrow reading it implies that moral judgments-*E* express only beliefs, while on a broad reading it implies that they express at least states of belief and perhaps other states as well In line with the resolution given earlier, genuine inconsistency obtains only with the narrow reading, and with this reading there is the anomaly that cognitivism, which is supposed to be represented in (1#), is only the contrary of noncognitivism, when the latter is put semantically as the view that moral judgments-*E* express moral attitudes and sentiment but not (true or false) beliefs To get (1#) and its denial to match up exactly with cognitivism and noncognitivism, we need the broad reading of (1#) and then the inconsistency vanishes Of course, there is a still broader reading of (1#), corresponding to (1*), namely that moral judgments-*E* *normally* express at least states of belief, to allow for the second kind of transition case, and again there is no inconsistent triad

But can there be a dual theory of moral judgments-*E*? That would require a dual theory of semantic content and would that be at all plausible? Rockney Jacobsen argues effectively against semantic cognitivism and for expressive pluralism (to use his terms) Only the former presents a problem for a dual theory of moral judgment-*E* Semantic cognitivism, in his words, says that "any utterance of a truth-assessable sentence which means that *p* is an expression of a belief that *p*"[50] This thesis has the unfortunate implication that when I say "I will arrive early" I must be expressing a belief rather than say a promise or a resolution or even an exclamation of surprise on discovering that I will arrive early Jacobsen writes

> *What* I promise, predict, or resolve is, in each case, that I will arrive early, and that would not be so unless the different illocutionary utterances had, in each case, the same meaning Indeed, it is only because meaning is autonomous from force that we can grasp the meaning of an utterance without knowing its force the query "Was that a promise, or a prediction?" betrays no lack of *semantic* knowledge (*ibid*, pp 133–34)

Truth value, he concludes, is independent of illocutionary force His second premise is "that the type of mental state expressed by an utterance is tied to, and varies with, its illocutionary force" (*ibid*,

[50] Jacobson, "Semantic Character and Expressive Content," *Philosophical Papers*, xxvi (1997) 129–46, at p 129

p 134) This premise is eminently plausible, assuming that the mental state need not always vary It follows that expressive content, that is, what mental state is expressed by an utterance, can vary without any change in meaning or truth aptness [51] For example, whether an utterance expresses surprise or a promise or a resolution or belief is not dependent on whether it is true or false Thus, Jacobsen concludes, semantic cognitivism is false

This argument appears to me to be sound Its conclusion bears directly on the present discussion, as Jacobsen is aware, since moral cognitivists have held that moral claims must be expressions of belief rather than attitude or desire, given that they are true or false, while moral expressivists have held that moral claims are expressions of attitude or desire and hence not true or false [52] Hence, both positions assume that any utterance of a truth-assessable sentence which means that *p* is an expression of a belief that p rather than attitude or desire That is, they assume semantic cognitivism, which is an untenable doctrine for the reasons just given Of course, Jacobsen's argument does not lead directly to the hybrid theory that I am advocating, but in an important footnote he allows it

> Another matter I cannot explore here my argument shows only that we are permitted, and often required, to count truth-apt judgments as expressions of feeling or desire, nothing in the argument shows that moral judgments do not *also* express beliefs It thus remains possible that my moral judgment expresses *both* a feeling of disapproval *and* my belief that what you are doing is wrong (*op cit*, p 145)

I believe that he is exactly right The point of immediate relevance, however, is that his argument clearly defeats the objection that dual semantic roles are impossible, since that objection is based on semantic cognitivism There exists no in principle objection to utterances functioning both to express belief and to express a different illocutionary force, such as commitment to behave in accord with moral norms, or to express emotions guided by those norms

VII WHAT IS MORAL JUDGMENT?

Moral judgment, whether judgment is taken psychologically or linguistically, is normally multifaceted involving intrinsic motivation, belief, and emotion A judgment that an action is morally wrong, for

[51] Copp works out this view in detail for the case of sentences expressing moral claims in his "Realist-Expressivism"

[52] The ecumenical versions of cognitivism and expressivism do not have this implication, but they offer no basis for the objection against the hybrid theory considered here

example, is normally (1) motivation of non-negligible force (or the expression of it) to treat the act as wrong just in virtue of certain perceived features, (2) belief (or expression of it) that the act is wrong, and (3) emotion (or the expression of it) that reinforces the motivation and the belief through its cognitive role in unifying these elements

The qualification 'normally' indicates that these elements are not always found together In the transition cases the elements can pull apart at least temporarily, then there is not one unified judgment but two different moral judgments each of which exemplify only some of the elements Each is nevertheless a moral judgment The theory describes the normal case and situates it in the context of stable moral norms Change in moral judgment does occur, however, and the theory explains how change in moral motivation and moral belief is possible and the cognitive role that emotion plays in it Another virtue of the theory is that it explains this possibility in a manner that resolves a central paradox regarding moral belief, feeling, and motivation and avoids the mistakes of the alternative realist theories of moral judgment

RICHMOND CAMPBELL

Dalhousie University

SELF-BIAS, TIME-BIAS, AND THE METAPHYSICS OF SELF AND TIME*

It is common to have a slightly exaggerated sense of the significance of your own joys and miseries, but grand self-importance is rare Louis XIV was grandly self-important He believed that, when he consumed too much foie gras, France suffered gastric pain When he took satisfaction from the construction of a new fountain in the grounds of Versailles, that feeling would settle over his natural kingdom—from the frigid, poxy docks of Brest to the steaming, cholera-ridden slums of Marseilles For Louis, self-indulgence was a national mission

> When we have the state in mind, we are working for ourselves The welfare of the one creates the glory of the other When the former is happy, lofty and powerful, he who is the cause of it has glory too and consequently should enjoy more than his subjects with regard to himself and to them everything in life that is most pleasant [1]

We, the undistinguished masses, are less fortunate than Louis As one of the masses I know that I have a set of unique qualities—a unique height, a unique weight, a unique time and place of origin and so forth—but these give me scarce grounds for thinking that I occupy a special place in the larger scheme of things, that my pleasures and pains are more significant than anybody else's I am not the physical embodiment of a national spirit I have not been chosen by God as his special representative on Earth I am hardly stronger, faster, or more delicate than the next person And there are seven billion next people

This means that, although I may be no less chronically attentive to my own comfort than Louis was, my self-attention is internally problematic in a way that his was not Louis took his pleasures and pains to be especially important, so when he deliberated he discovered a neat harmony between *egocentric-hedonistic considerations* (of the form 'overall, I will suffer less if I do __ rather than __') and *considerations of the greater good* (of the form 'overall, things will be better if I do __ rather than __') When I deliberate I discover no such harmony I am continually faced with situations in which these considerations conflict,

* Many thanks to Anthony Appiah, Paul Benacerraf, Karen Bennet, Akeel Bilgrami, Ned Hall, Simon Keller, Rae Langton, Gideon Rosen, and in particular Mark Johnston for helping me to develop the ideas here

[1] Louis XIV, *Mémoires for the Instruction of the Dauphin*, Paul Sonnino, trans and ed (New York The Free Press, 1970)

0022-362X/07/0407/350–73 © 2007 The Journal of Philosophy, Inc

situations in which I can make the world better at a cost to my comfort, and myself more comfortable at a cost to the world

That may seem a very obvious point Of course, people tend to be disproportionately preoccupied with their own comfort Of course, to be so preoccupied is not to be preoccupied with the greater good

Can it be denied?

I SELF-BIAS

To get a grip on this question, let us consider the position of a *peacemaker,* a person whose desire for the kind of psychological harmony enjoyed by Louis leads her to be resolutely committed to the view that, for the most part, concern for one's own comfort and concern for the greater good align The peacemaker notes that although most of us are not *extreme egocentric-hedonists,* caring only about our own pleasures and pains, we are at least *mild egocentric-hedonists*—all other things being equal, we prefer that pain befall others rather than ourselves, and pleasure befall ourselves rather than others The peacemaker then claims

(Harmony) Whenever a mild egocentric-hedonist favors a scenario in which she suffers less, she thereby favors a simply better maximal state of affairs

I have used some quasi-technical terms here, and they need to be glossed What is it to for one thing to be 'simply better' than another? Let us just say what it is not—not better in some qualified or three-way relational sense of the word, not better for me, or better for some purpose, or better in relation to a particular set of interests, but better period What is a 'state of affairs'? Think of a state of affairs as a way for things to be So the way things in Yosemite National Park are is a state of affairs The way things in the Russian economy are is a state of affairs What is a 'maximal state of affairs'? Well, some states of affairs can be parts of others—the way things in my laundry cupboard are is a part of the way smelly things in the Boston area are A maximal state of affairs is one that is not a proper part of any state of affairs It is a way for everything to be Finally, what is it to 'favor a scenario'? I mean the notion of 'favoring' to be understood quite broadly one way to favor a scenario or state of affairs is to bring it about, another is just to desire that it obtain So, according to (Harmony), it is not just that the actions of a mild egocentric-hedonist bring about the good (whenever she brings about a situation in which she suffers less, she brings about a better maximal state of affairs), it is also that the desires of a mild egocentric-hedonist align with the good (whenever she wants a situation in which she suffers less to come about, she wants a better maximal state of affairs to come about)

Is the peacemaker's claim credible? Well, three seemingly insurmountable obstacles stand in the way of my believing it

I A The Grounding Problem First, if I am to believe it, I must take it to be true of *me* that when I enhance my own comfort at the expense of other people, I am making for a better overall state of affairs So I must adopt a picture of the world in which my comfort makes an especially weighty contribution to the value simpliciter of a state of affairs Call this the *Grounding Problem* my picture of who I am, and of how I fit into the world, must somehow give me grounds for thinking that my pains and pleasures have an especially important place in the larger scheme of things

But how could this be? I am not the Sun King I am mundane in all measurable respects So perhaps my only hope is to believe that I am extraordinary in some immeasurable respect, that I am metaphysically unique in some way that bears upon the intrinsic value of my suffering For example

The Solitary Utility Monster

> My physiology is not unusual in any measurable way—I wince when punched, smile when pleased, and frown when saddened, just like anyone else But there is one important, hidden difference between me and other people my experiences are, qualitatively, far more intense than theirs If their experiences are like watercolors by Joseph Turner, mine are like oil paintings by Francis Bacon Drinking tea feels to me just the way that taking morphine feels to them A touch of indigestion feels to me just the way that being disemboweled feels to them

If this bleak picture were accurate then perhaps I should be a mild egocentric-hedonist For surely, one could argue, what is intrinsically good or bad about pleasure or pain is its phenomenal aspect Hit your hand hard against your desk What is bad about pain is that feeling Run your fingers gently across your scalp What is good about pleasure is that feeling And there are more of those sorts of feelings associated with my experiences than with anybody else's

But the picture is, of course, ridiculous Serious philosophers do not believe that they are metaphysically unique At best they use the idea as a foil—the *problem* of the "problem of other minds" is almost always taken to be the problem of how one can know that other people's mental states are just like one's own, very rarely taken to be the problem of whether other people's mental states are just like one's own And to my knowledge, no serious philosopher has ever even considered the view that I, Caspar Hare, am metaphysically unique It has not even been in the ballpark[1]

So this does not seem like a very promising strategy If I am to solve the Grounding Problem, I must do so without committing myself to a wildly implausible metaphysical picture And that appears impossible

I B The Generalization Problem Furthermore, if I am to argue, quite generally, that there is no conflict between the considerations that move a mild egocentric-hedonist and considerations of the greater good, it is not enough just to show that there is no conflict in my own case I must show that any mild egocentric-hedonist, in favoring scenarios in which she suffers less, is favoring better maximal states of affairs And it is very difficult to see how this can be so Call this the *Generalization Problem*

To see the difficulty, imagine a situation in which my interests are at odds with someone else's For example

Competing for a Scarce Resource

Jane and I (CJH) are competing for a scarce resource Jane will suffer less if she gets it I will suffer less if I get it

In this situation there seem to be two possible ways for things to go

Possibility (1)	CJH gets the resource and is happy	Jane misses out and is miserable
Possibility (2)	CJH misses out and is miserable	Jane gets the resource and is happy

Being a mild egocentric-hedonist, I favor (1) over (2) Being a mild egocentric-hedonist, Jane favors (2) over (1) But, if the notion of 'better simpliciter' is to remain coherent, it cannot both be better simpliciter that the state of affairs represented by possibility (1) obtain and better simpliciter that the state of affairs represented by possibility (2) obtain So surely in this case at least one of us, in favoring the state of affairs in which we suffer less, is not favoring a simply better maximal state of affairs [2]

I C The Problem of Irreducibly Egocentric Preferences The third problem arises from the fact that, for me, caring about CJH's comfort and caring about *my* comfort do not always amount to the same thing Given that I am broadly egocentric, and that I believe myself to be

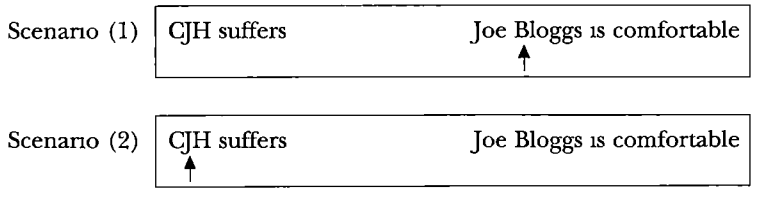
[2] This is the thought behind the famous argument against egoism in Section 59 of G E Moore's *Principia Ethica* (New York Cambridge, 1903) "The only reason I can have for aiming at 'my own good' is that it is good absolutely that what I so call should belong to me—good absolutely that I should have something, which, if I have it, others cannot have What Egoism holds, therefore, is that each man's happiness is the sole good—that a number of different things are each of them the only good thing there is—an absolute contradiction!"

CJH, I want to promote my comfort and the comfort of CJH But, given that I am broadly egocentric, the moment I ceased to believe that I was CJH, I would cease to care about *him* and continue to care about *me* There are imaginable situations in which this might happen

After the Train-Crash

I wake up in hospital, achy and bewildered, unable to remember who or where I am I try to move, and find that my body is swathed in rigid plaster and my head is locked in a brace I call for help and receive no reply But, happily, some kind nurse has placed a television directly in front of me From it I learn that there has been a terrible train accident, that only two survivors, CJH and Joe Bloggs, have been pried from the wreckage, and that both have been taken to hospital and placed in full-body plaster casts, from which they are watching this very program on television "Interesting!" I think, "I now have a vivid picture of what's going on in this hospital, and I know a lot about CJH and Joe Bloggs, although I still don't know which of them I am " Next the television tells me that one of the two is scheduled to have a long and painful operation in a few hours time "Interesting!" I think, "I hope that's not *me*"

My first concern in this situation, is for me, not CJH If I were to discover that CJH was to be the unfortunate subject of this operation, I would be neither happy nor unhappy, because two importantly different possible scenarios would remain open, scenarios that we can represent like this (with arrows to represent who I am)



I very much want the first scenario to obtain I want to be Joe Bloggs, the person who will not be suffering

But surely this is a clear example of a person favoring a scenario in which he is better off without thereby favoring a better maximal state of affairs For, on the standard account of the content of "self-involving" propositional attitudes,[3] when I hope, desire, believe, or discover that scenario (1) obtains, I do not hope, desire, believe, or discover that the

[3] Classic variants of which are in Hector-Neri Castañeda, "On the Logic of Attributions of Self-Knowledge to Others," this JOURNAL, LXV, 15 (August 8, 1968) 439–56, and John Perry, "The Problem of the Essential Indexical," *Noûs*, XIII (1979) 3–21, and David Lewis, "Attitudes De Dicto and De Se," *The Philosophical Review*, LXXXVIII (1979) 513–43

world is a certain way After the operation I already know everything relevant about the way the world is, about the nature of the maximal state of affairs—I know who will suffer, when they will suffer, and how much they will suffer My preference for (1) over (2) is not a preference that the world be one way rather than another, but rather a preference that I be one way rather than another, it is, to use some jargon, an irreducibly *de se* preference, a preference with no *de dicto* content

In light of the Grounding, Generalization, and Irreducibly Egocentric Preferences problems, it appears that (Harmony) must be rejected and the peacemaker's project must be abandoned The pressing question for ethicists to address is not whether the considerations that move a mild egocentric-hedonist misalign with considerations of the greater good, but how we should respond when they do And, sure enough, ethicists have been responding to some form or other of this question for a very long time

That may seem like a compelling argument, but it relies upon some substantive and questionable assumptions about the metaphysics of the self Perhaps the clearest way to bring this out is by looking at an analogous argument, to the conclusion that we cannot find harmony between considerations of the greater good and time-biased considerations

II TIME BIAS

Some people care not only about *what* things happen, but also about *when* things happen One way to care about the *when* as well as the *what* is to care about how events are ordered over the course of history—a history across which good and bad things are evenly sprinkled, for example, might seem preferable to one with good things clumped at one end, bad things clumped at the other Another way is to care more about what happens at some times than others—what happens on the first day of the year 2000, for example, might seem to matter more than what happens on the three hundred and twenty sixth day of the year 1994 Another way is to care about when things happen relative to the present moment Call someone who cares about when things happen relative to the present moment *time-biased*

Of the many ways in which one might be time-biased, the two that have received most attention from philosophers[4] are *hedonic bias toward the future* (all things considered, I prefer that pains be past rather than future, and pleasures be future rather than past), and *hedonic bias toward the near* (all things considered, I prefer that pains be in the distant future rather than the immediate future, and pleasures be in the

[4] Largely due to the discussion in chapter 8 of Derek Parfit, *Reasons and Persons* (New York Oxford, 1984)

immediate future rather than the distant future) Everybody seems to some extent vulnerable to these kinds of bias Would you not prefer to be walking out of your dentist's office, with the pain in your tooth subsiding, than to be walking in, with the bulk of the pain still to come? Would you not prefer the drill to be a week, rather than an hour away?

Let us restrict our attention to pain Someone who has these sorts of biases may give them a role in her practical reasoning by taking considerations of the form 'there is less future pain in __ than __' and 'there is less near-future pain in __ than __' to support favoring one scenario over another But obviously it would be hasty to take such considerations to provide the last word in practical deliberation, because sometimes they will misalign or conflict with considerations of the greater good Sometimes, when you desire, or make it the case that there is less future or near future pain, you do not desire, or make it the case that things overall are better

Why is this obvious? Once again it will be useful to consider a *peacemaker*, who wants to claim that time-biased considerations are in harmony with considerations of the greater good The peacemaker is committed to

(Harmony 2) Whenever a person favors a scenario in which there is less future, or near-future pain, she thereby favors a simply better maximal state of affairs

Can she really believe this? Three familiar obstacles stand in her way

II A The Grounding Problem First, she would have to adopt a metaphysical picture that gave her some grounds for thinking that future pains are in themselves worse than past pains, near-future pains in themselves worse than far-future pains Her picture would need somehow to support the idea that Joan of Arc's suffering on the pyre matters less than mine at my next workout Here is a picture that would work

Divine Novocaine

At the dawn of creation God faced a problem He would not allow himself physically to intervene in the course of history, for fear of compromising our freedom, but he wished to protect us from the evil consequences of that freedom So he created *Divine Novocaine*, a drug that dulls the qualitative aspect of suffering without inducing any physical changes in the sufferer In the past he has administered this drug liberally, to all sufferers, pure and fallen, but now (confronted with famine, war, and disease on a scale that even he could not have anticipated) he finds his stocks to be running low He will be forced to ration So he decides that he will administer none for a short while, and then gradually build up the doses, though never to their pre-2008 level

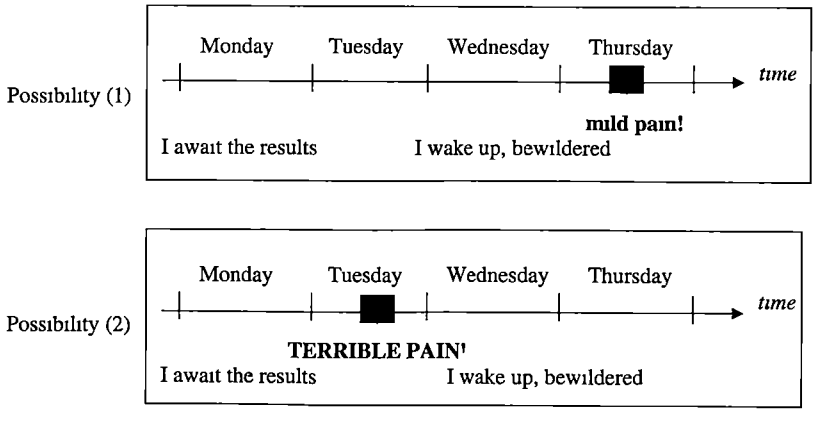
This would do the trick But it is mad On any sane view, Joan of Arc's suffering is no less intense and no less significant for being past

II B The Generalization Problem Second, it is not enough for our peacemaker to show that considerations of the greater good align with time-biased considerations *now,* in 2008 She needs to show that they *always* do So she needs to deal with cases where time-bias leads us to have different preferences at different times about the same states of affairs Here is an example of such a case, where the preferences are generated by bias toward the future [5]

Bad and Worse Operations

I must have an operation It is a safe operation, but it involves some nasty stuff some cutting, some scraping, some poking of optic fibers down blood vessels, and so forth On Monday doctors perform tests on me and declare that, depending on the results of these tests, I will either endure the operation under the influence of a local anesthetic (unpleasant, but not painful) on Thursday, or endure the operation under the influence of no anesthetic at all (terribly painful) on Tuesday On Monday I hope to have the unpleasant operation on Thursday But on Wednesday evening I wake up, bewildered, unable to remember if I have yet been operated upon, and, since I am biased towards the future, I hope to have had the terribly painful operation on Tuesday

Here the possible world-histories seem to be



[5] Are there similar such cases where the preferences are generated by bias toward the near? This will depend upon the rate at which you discount for temporal distance If you discount exponentially (for example, where i is the intensity of a future pain and t is the temporal distance, by taking its significance to you now to be given by $i/2n$) then your preferences will not change over time If you do not then they will See section 62 of Parfit, *Reasons and Persons*

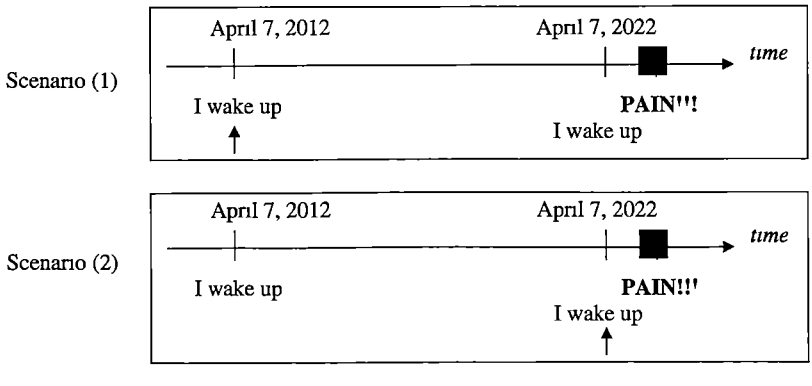
On Monday I prefer (1) to (2), while on Wednesday I prefer (2) to (1) But it cannot both be better simpliciter that the maximal state of affairs represented by possibility (1) obtain, and that the maximal state of affairs represented by possibility (2) obtain So on at least one of the two days considerations of the greater good do not align with future-biased considerations On at least one of the two days, in favoring a scenario in which there is less future-pain, I am not favoring a better maximal state of affairs

II C The Problem of Irreducibly Egocentric Preferences Third, the peacemaker must face cases where time-bias leads us to have irreducibly egocentric preferences Here is an example of such a case, where the preference is generated by hedonic bias toward the near[6]

Waiting For My Painful Operation

Early in my life, clever doctors find that I have a rare genetic condition that will most likely cause me to contract cancer of the x (substitute whatever organ you feel most anxious about for x) in late middle-age Happily, they can eliminate the risk by removing organ x from my body Unhappily, the necessary operation is agonizingly painful Knowing all this, I schedule such an operation for soon after April 8, 2022, my fiftieth birthday I am always perfectly sure that the operation will take place, that I will survive it, that it will be successful, and that it will hurt like anything Then, one night, I wake up from a fretful dream and look at the date on my malfunctioning alarm clock It says 'April 7, 2' Is it April 7, 2012 or April 7, 2022? For the moment I just don't know But I do know what I want Being biased toward the near, I want it to be 2012

In this case there are two possible scenarios I might be in (with arrows to represent what the time is now)



[6] Future-bias generates many similar preferences—for example, when I wake up, knowing that I have a painful operation on a particular date, but not knowing whether it is in the past or future

But these two scenarios do not represent different ways for the world to be, different maximal states of affairs My ignorance in this case is, in the jargon, irreducibly de se—I know everything relevant about the history of the world, but do not know which person-slice *I* am So, in favoring (1) over (2), I am not favoring one maximal state of affairs over another So, in favoring the scenario in which there is less near-future pain, I am not favoring a better maximal state of affairs

As before, these three problems may seem to be insurmountable, so it may seem that the peacemaker must give up, and concede that no plausible metaphysical picture grounds general time-bias Time-biased considerations do not always align with considerations of the greater good

III TIME-BIAS AND THE METAPHYSICS OF TIME

The argument above is structurally just like the argument from section I, the argument that egocentric considerations do not always align with considerations of the greater good Is there anything wrong with it? Well, this time around we have at least some reason to suspect that there might be something wrong with it, because its conclusion is controversial Certainly, many influential philosophers (including Henry Sidgwick[7] and John Rawls[8]) have had a low view of time-bias, arguing that it is petty, unenlightened, or, in the extreme, irrational, but others (including Jeremy Bentham[9]) have thought it to be perfectly rational, and furthermore a bias toward what is *better* It is better that, for example, suffering be in the past rather than the future

One way to accommodate this intuition would be to take issue with my characterization of the sorts of states of affairs whose intrinsic value gives rise to considerations of the greater good I took them to be maximal states of affairs, but one could perhaps take them to be proper parts of maximal states of affairs In favoring better futures I may, then, be favoring intrinsically better "states of affairs" of the relevant kind (Lars Bergstrom once endorsed this view, arguing that, for consequentialists, "the future is more important than the total state of the world "[10]) Or, for a quite general solution to problems of agent and time-bias, one could stipulate that the relevant "states

[7] See Sidgwick, *The Methods of Ethics* (London Macmillan, 1907), pp 381, 382

[8] See section 45 of Rawls, *A Theory of Justice* (Cambridge Harvard, 1971)

[9] that is, on one reading of chapter IV of Bentham, *An Introduction to the Principles of Morals and Legislation* (1789) Bentham's views may be somewhat underdetermined by the text

[10] Bergstrom, *The Alternatives and Consequences of Actions* (Stockholm University Press, 1966), p 125

of affairs" are <*World,Agent, Time*> triples, the "centered worlds" that figure prominently in decision theory It may be the case that (for example), in always favoring better futures, I am always favoring better "states of affairs" in this sense, because it may be that whenever $<w_1, a_1, t_1>$ and $<w_2, a_2, t_2>$ are such that there is less suffering in w_1 after t_1 than in w_2 after t_2, then $<w_1, a_1, t_1>$ is simply better than $<w_2, a_2, t_2>$

But there is something fishy about this suggestion For one thing, notice that there is negligible difference between saying that the triple $<w, a_1, t_1>$ is simply better than the triple $<w, a_2, t_2>$, and saying that world w is better for, or in relation to, a_1 at t_1 than for, or in relation to a_2 at t_2 For another thing, it mischaracterizes the nature of considerations in question Considerations of the greater good gain their force from the thought that what matters is the way that everything is, not the way that everything is relative to a time, place, or person But the triples $<w, a_1, t_1>$ and $<w, a_2, t_2>$ do not represent different ways for everything to be

Another, much better way to accommodate the intuition is to question whether facts about what is happening now do not somehow enter into the way that everything is This involves making some substantive assumptions about the metaphysics of time

III A Four-Dimensionalism One vexed question in the metaphysics of time concerns the ontological status of past, present, and future entities (moments, events, objects) Does Julius Caesar exist? Does the nineteth President of the United States exist?

Call someone who believes that past and future entities (moments, events, objects, and so on) exist an *eternalist* Some eternalists are *ersatz realists* about the past and future, they believe that present things are concrete, past and future things abstract Others believe that past, present, and future things are equally real, equally concrete "The world consists of a four-dimensionally extended space-time manifold," these people say, "the past, present and future are ontologically on a par Just as a thing is no less real for being to my left or right, so a thing is no less real for coming before or after me " Call such people *block universe theorists*

Another vexed question concerns the status of tensed properties What is it for something to be past, present, or future? Some block universe theorists think that there are no monadic tensed properties, only relational tensed properties—the death of King Harold is not monadically past, present, or future, it is past relative to the death of Ann Boleyn, future relative to the death of Jesus, and present relative to William's invasion of England "All things are on a par with respect to tense," these people say, "all things are future relative to things that come before them, present relative to things with which they coexist,

and past relative to things that come after them No moment has the privileged status of being the present one " Call such people *four-dimensionalists* [11]

Now, if we accept four-dimensionalism, then the Grounding, Generalization, and Egocentric Preferences problems really are insoluble With respect to grounding, past pains are just as real as future pains, so we have no real grounds for thinking them less important simpliciter With respect to generalization, in the *Bad and Worse Operations* case, my preferences on Monday and Wednesday are indeed inconsistent On Monday, I favor the one space-time manifold over the other, while on Wednesday, I favor the other over the one, so I cannot, on both days, be favoring better ways for everything to be With respect to irreducibly egocentric preferences, in the *Waiting for My Painful Operation* case, my desire that pain be in the far future rather than the near future does indeed have no de dicto content I am not favoring one space-time manifold over another, so I am not favoring a better over a worse way for everything to be

So if we accept four-dimensionalism then the argument goes through—considerations of the greater good do not always align with time-biased considerations But what if we deny it? What if we insist that there are monadic tensed properties, that tense is built into the way the world is?

III B Alternatives to Four-Dimensionalism There are several ways to accommodate this basic thought Advocates of *hyper-kenesis*[12] retain the block universe ontology, retain the idea that the world is a four-dimensionally extended space-time manifold, but say that one moment has the interesting and unique property of being the present one Which moment? Well, that changes as time goes by (think of a spotlight moving remorselessly along the block) Others imagine that the past exists but the future does not The present is the outermost skin of a block that expands as time goes by [13] Others imagine that the future exists but the past does not The present is the outermost skin

[11] This term has been used by several different philosophers in several different ways—sometimes to mean the block universe picture, sometimes (as here) to mean the block universe picture conjoined with the view that tensed properties are relational, sometimes to mean the block universe picture conjoined with the view that tensed properties are relational conjoined with perdurantism about persistence over time So, for maximum precision, perhaps I should use a different term, but I find the one that suggests itself—'the uncentered block theory'—unacceptably clunky

[12] See, for example, George Schlesinger, *Timely Topics* (New York St Martin's, 1994)

[13] This view may have origins in Aristotle A contemporary version of the view is proposed by Michael Tooley, in *Time, Tense and Causation* (New York Oxford, 1997)

of a block that contracts as time goes by [14] Others imagine that the future branches off in many directions while the past remains fixed The present is the first point at which multiple branches split off from the bare tree-trunk of the past [15] Presentists,[16] meanwhile, hold that only present objects, events, moments exist (along, perhaps, with timeless things like gods and numbers) There are no past or future things The best we can say is that it used to be the case that they existed, or it will be the case that they exist

Each of these is a theory about the nature and extent of all that there is, where 'all' is understood in an unrestricted sense—a theory about what maximal states of affairs are like And, according to each of them, it is part of the nature of maximal states of affairs that present things are different from past and future things So, according to each of them, the Generalization and Egocentric Preferences problems do not arise With respect to generalization in the *Bad and Worse Operations* case the maximal states of affairs that I discriminate between on Monday (one in which mild pain is in the far future and one in which terrible pain is in the near future) are not the same as the maximal states of affairs that I discriminate between on Wednesday (one in which terrible pain is in the past and one in which mild pain is in the near future), so it is quite possible that, on both days I am favoring a simply better maximal state of affairs With respect to irreducibly egocentric preferences in the *Waiting for My Painful Operation* case, the two scenarios that seem open to me really do represent two different maximal states of affairs (in the one case a maximal state of affairs in which pain is in the far future, in the other a maximal state of affairs in which pain is in the near future), so it is quite possible that, in favoring one scenario over another, I am favoring a simply better maximal state of affairs

So if a dedicated peacemaker rejects four-dimensionalism, and accepts that there are monadic tensed properties, she can solve the Generalization and Irreducibly Egocentric Preferences problems by stipulating that, all other things being equal, states of affairs in which

[14] Though many people have imagined this, I do not know of anyone who has endorsed it Surprising, perhaps, because it seems uniquely well qualified to explain why the future matters and the past does not

[15] A view developed in Storrs McCall, *A Model of the Universe* (New York Oxford, 1994)

[16] The first and most famous presentist was Saint Augustine For more up-to-date expositions of the view, see John Bigelow, "Presentism and Properties," in James Tomberlin, ed , *Philosophical Perspectives 10* (Cambridge Blackwell, 1996), pp 35–52, and Dean Zimmerman, "Temporary Intrinsics and Presentism," in Peter van Inwagen and Zimmerman, eds , *Metaphysics The Big Questions* (Malden, MA Blackwell 1998), pp 206–19

pain will occur are worse than states of affairs in which pain has occurred, and states of affairs in which pain will happen soon are worse than states of affairs in which pain will happen a long time from now

Would this solve the Grounding Problem? Would the metaphysics somehow make it seem plausible that pain with the monadic property of being in the future is in itself more significant than pain with the monadic property of being in the past? Well, unless our peacemaker adopts the "shrinking block" view, the metaphysics will not explain why past pains matter less than future pains But maybe no explanation is needed If the peacemaker has a strong (though perhaps defeasible) conviction that future pain is intrinsically worse than past pain, the metaphysics does nothing to undermine her conviction

IV SELF-BIAS AND THE METAPHYSICS OF THE SELF

We can make peace between considerations of the greater good and time-biased considerations by adopting an appropriate metaphysical picture Can we perform an analogous trick for egocentric considerations? What sort of picture would allow us to say that, when we favor scenarios in which we are better off, we favor simply better maximal states of affairs? Well, at a minimum the picture would have to imply that, in the *After the Train Crash* case, where I have an irreducibly egocentric preference, the two possibilities represent different ways for the world to be But since these "possibilities" differ only with respect to which of the injured parties is *me*, the picture would have to imply that this property, *being-me*, somehow enters into states of affairs

The analogy to four-dimensionalism and its denial should be clear According to a widely accepted view of the self, if *being-me* and *being-other* are properties at all, then they are relational properties CJH is me relative to CJH, other relative to Stalin Stalin is me relative to Stalin, other relative to Lenin And everybody is, in the relevant sense, on a par with respect to these properties—we are all me relative to ourselves, and other relative to everybody else But according to the rather unorthodox view we would have to take, it would have to be the case that in any given maximal state of affairs, one and only one person has the nonrelational property of being-me There is a unique *I* at the center (so to speak) of all that exists *After the Train Crash*, I am comparing maximal states of affairs in which the thing with the monadic property of being-me is in pain, with maximal states of affairs in which the thing with the monadic property of being-me is free of pain

Will this not turn out to be just another theory of the kind we considered in section I, another theory that says I am metaphysically unique? Well, yes, but if it is to solve the Generalization Problem then

it needs to be considerably more subtle than those were Let us say that I am CJH, and Jane and I are each contemplating a state of affairs in which CJH suffers and Jane prospers Our theory will need to explain why, although the state of affairs I am considering is one in which the thing with the monadic property of being-me suffers, the state of affairs Jane is considering is not And it will need to explain why, when Jane says 'I am going to prosper', the proposition expressed by her sentence is true

It may be helpful to see how a someone who believes in monadic tensed properties (henceforth a *monadic tenser*[17]) deals with the analogous "problem " Let us say that it is presently noon, and Joe at noon and Joe at 11A M are both contemplating a situation in which Joe at noon suffers, and Joe at 11A M prospers The monadic tenser must explain why the state of affairs that Joe at 11A M contemplates is one in which Joe presently prospers, not one in which he prospered one hour ago And she must explain why, when Joe at 11A M says 'Joe is not suffering', the proposition expressed by his sentence is true

Obviously this poses no real difficulties for the monadic tenser If it is presently noon, and Joe is suffering, and Joe at 11A M was thinking about this, then the situation that Joe at 11A M *was* contemplating *was* one in which Joe *was not* suffering And *it was the case that,* when Joe said 'Joe is not suffering', he said something true

The point is that the monadic tenser has use of operators—*it will be the case that* and *it was the case that*[18] These operators enable her to make sense of the states of affairs that past and future people consider, and to make sense of past and future tensed utterances And they may also allow her to take the some of the counter-intuitive sting out of her metaphysical commitments For example, a presentist can say that, although Anne Boleyn does not exist, she is not like Santa Claus, because *it was the case that* (Ann Boleyn exists) Although my first granddaughter does not exist, she is not like the boogeyman, because *it will be the case that* (my first granddaughter exists) In general, for any thing that the four-dimensionalist believes to exist, the presentist can say that either it exists, or it will exist, or it has existed The present is special, then, but not quite as special as it might at first have appeared

The theory we are groping towards is going to need operators that, in a similar kind of way, allow us to make sense of self-referential

[17] Such people are sometimes called *serious tensers* But that seems prejudicial to me

[18] Monadic tensers typically also avail themselves of so-called "metric" operators for any *n*, the monadic tenser has operators *in n seconds it will be the case that* and *n seconds ago it was the case that* But we will not need these for the moment

sentences uttered by other people, and of the "possibilities" that other people consider Which operators we use will depend on the details of the theory we adopt I will outline a theory,[19] call it *Egocentric Presentism*, which seems to me attractive and simple

IV A An Introduction to Egocentric Presentism The best way to introduce this theory is with a story Imagine that, in a fit of Cartesian pique, I throw away all of my cherished beliefs about how the world is, and about how I fit into it I no longer accept that the earth is round, that there are material objects, that I am CJH, or, for that matter, that I am anything at all In this epistemically emaciated state I retire to a secluded room and attempt to build up a world-view from there A series of insights strike me

Insight 1 There are some things a painting of Saint George and the Dragon, a telephone, a diary, a facial itch Their nature remains obscure Perhaps the painting is not really a painting Perhaps the telephone is a toy Perhaps the diary is a bundle of sense data Perhaps the itch is imaginary All that can be said for sure is that these things, whatever they are, reveal themselves at this, the first stage of the Cartesian exercise They are *present*[20]

Insight 2 The present things are perceptual objects of a sentient being, CJH CJH sees the telephone, painting, and diary CJH feels the itch

Insight 3 There are many, many sentient beings other than CJH, but their perceptual objects are not present So as to talk about this interesting feature of CJH in an economical way, let us introduce some terms Let 'present experiences' be short-hand for 'experiences whose objects are present' Let 'I' (in the nominative, 'me' in the accusative) be short-hand for the definite description 'the one with present experiences'

[19] I will not place the theory in any historical context here, but it picks up on themes from the work of a great many philosophers It could, for example, be presented as the solution to a problem that troubled Ludwig Wittgenstein, the problem of how to retain the basic Humean idea about self-knowledge (that the truth with which I am confronted, when I direct my attention towards my own experience, is not *that I am in pain* but *that there is pain*, not *that I see red* but *that there is a sensation of red*) without being forced to be a solipsist about other minds See Wittgenstein, *Tractatus Logico-Philosophicus*, C K Ogden, trans (Mineola, NY Dover, 1922), §5 557–§5 641, and Wittgenstein, *Philosophical Investigations*, G E M Anscombe, trans (Upper Saddle River, NJ Prentice Hall, 1953), §402–§420

[20] This is a term of art, and I introduce it guardedly The advantages of using the word 'present' in this context are that it supports both monadic and relational readings, and that, by doing so, I emphasize the structural similarities between this view and the analogous views in the metaphysics of time

Insight 4 Although I, CJH, am unique in having present experiences, I find that I can imagine other sentient creatures being unique in this respect I can imagine, for example, that I am Michael Schumacher (that the one with present experiences is Michael Schumacher) This involves imagining that Michael Schumacher's perceptual objects—heat, a smell of foam and latex, the wail of a ten-cylinder engine—are present It involves imagining hoardings buzzing past me, the one with present experiences, Michael Schumacher, at two hundred miles per hour

Insight 5 Such imaginings lead me to suspect that I am less unique than I may have thought Michael Schumacher's experiences are not present, so things are not as I imagined them to be, when I imagined being him But he is conscious, he has a point of view, and from his point of view things *are* as I imagined them to be—the heat, the foamy smell, and the wail of the engine *are* present From Michael Schumacher's point of view, Michael Schumacher's experiences are present

But let us be careful here In saying that Michael Schumacher is conscious I am not taking presence to be a relational property—saying that his experiences are present *relative to his point of view,* while mine are present *relative to my point of view* Presence is a monadic property that my experiences have and his do not The construction 'from his point of view' is really an operator, the semantics for which might be given in the following way

IV B Semantics for a Logic of Points of View Say that a *subject world* (henceforth an *S*-world) is a world in which there are functionally sentient creatures, the experiences of one and only one of which have the monadic property of *being-present*

At any such world, a set of *atomic* propositions hold true Think of these as propositions having to do with the way things are, physically speaking, and propositions having to do with where the property of being present is instantiated So, for example, at $S_{\text{HenryKissinger}}$, a world physically identical to our own, but in which the experiences of Henry Kissinger are present, the following atomic propositions hold true

'The sun orbits the moon'

'There are over ten billion functionally sentient animals'

'The person with present experiences was formerly Secretary of State'

'CJH's experiences are absent' and so on

Now, let a *system* of *S*-worlds be a set of physically identical *S*-worlds such that for any functionally sentient creature in an *S*-world in the set, there is an *S*-world in the set in which that very creature has

present experiences And say that for any system of S-worlds there is a reflexive two-place access relation, the a-relation, defined over pairs of S-worlds in the system

Here, for the sake of having an example to work with, is one system of S-worlds

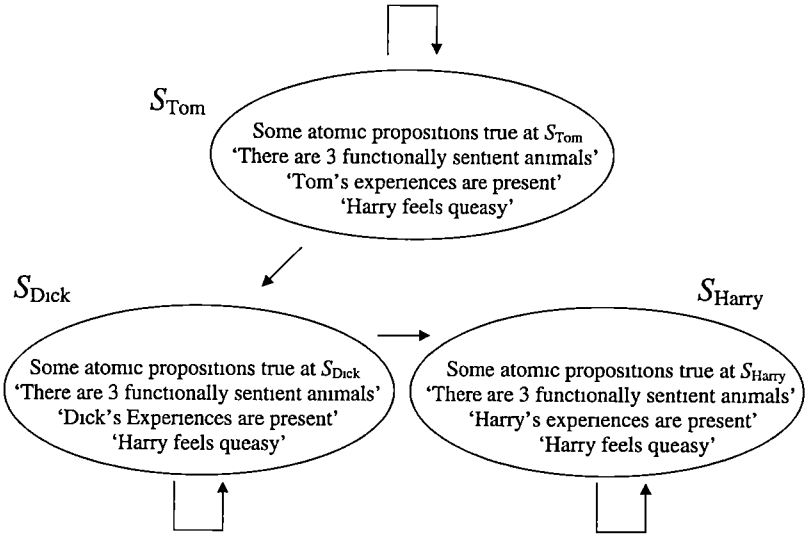

S_{Tom}
Some atomic propositions true at S_{Tom}
'There are 3 functionally sentient animals'
'Tom's experiences are present'
'Harry feels queasy'
S_{Dick}
Some atomic propositions true at S_{Dick}
'There are 3 functionally sentient animals'
'Dick's Experiences are present'
'Harry feels queasy'
S_{Harry}
Some atomic propositions true at S_{Harry}
'There are 3 functionally sentient animals'
'Harry's experiences are present'
'Harry feels queasy'

Figure 1

(The ovals represent S-worlds The arrows represent access relations between S-worlds $S_i \rightarrow S_j$ means S_i is a-related to S_j) So, in this system there are three S-worlds, S_{Tom}, S_{Dick}, and S_{Harry} They are all a-related to themselves, S_{Tom} is a-related to S_{Dick} (though not vice-versa), and S_{Dick} is a-related to S_{Harry}, (though not vice-versa)

Now we give truth conditions for sentences containing quantified *point of view* operators in the following way

Definition 1 From some point of view, from every point of view

$\ulcorner$*From some point of view* $(p)\urcorner$ is true at S_K iff for some S_J a-related to S_K, $\ulcorner p \urcorner$ is true at S_J

$\ulcorner$*From every point of view* $(p)\urcorner$ is true at S_K iff for every S_J a-related to S_K, $\ulcorner p \urcorner$ is true at S_J

So, for example, in the above system the proposition '*From some point of view* (there are present queasy experiences)' is true at S_{Dick}, because S_{Dick} is a-related to S_{Harry} and 'there are present queasy experiences'

is true at S_{Harry} But it is false at S_{Tom}, because S_{Tom} is not a-related to S_{Harry}, or to any other S-world at which 'there are present queasy experiences' is true

And we give truth conditions for sentences containing operators of the form *from H's point of view* with the following definition schema

Definition 2 From H's point of view

For any functionally sentient being H, $\ulcorner$*From H's point of view* $(p)\urcorner$ is true at S_K
iff
for some S_J a-related to S_K, $\ulcorner H$ has present experiences, and $p\urcorner$ is true at S_J

So in the above system, for example, '*from Harry's point of view* (Harry's experiences are present and Harry is queasy)' is true at S_{Dick}

Finally we stipulate that

Definition 3 Consciousness

$\ulcorner H$ is conscious$\urcorner$ is true at S_K
iff
$\ulcorner$*From H's point of view* (H's experiences are present)$\urcorner$ is true at S_K

So in the above system, for example, 'Harry is conscious' is true at S_{Dick} but not at S_{Tom}, because S_{Dick} is a-related to a world in which Harry's experiences are present, but S_{Tom} is not But '*from Dick's point of view* (Harry is conscious)' is true at S_{Tom}, because S_{Tom} is a-related to S_{Dick}

Here is a different system

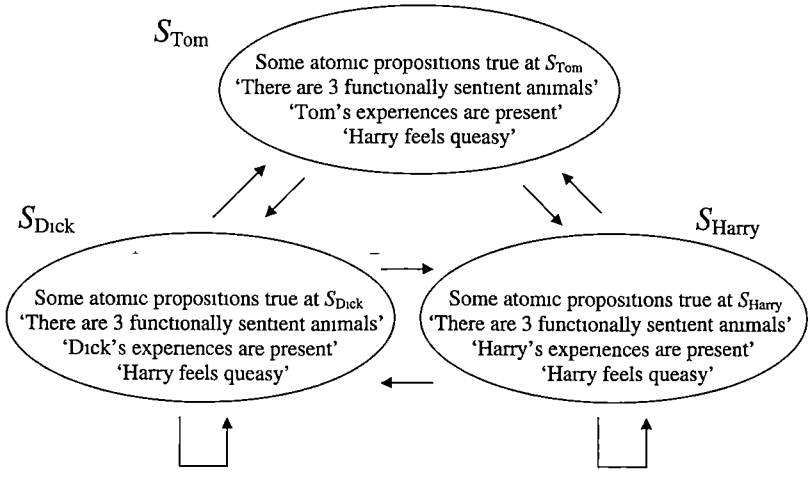

S_{Tom}
Some atomic propositions true at S_{Tom}
'There are 3 functionally sentient animals'
'Tom's experiences are present'
'Harry feels queasy'
S_{Dick}
S_{Harry}
Some atomic propositions true at S_{Dick}
'There are 3 functionally sentient animals'
'Dick's experiences are present'
'Harry feels queasy'
Some atomic propositions true at S_{Harry}
'There are 3 functionally sentient animals'
'Harry's experiences are present'
'Harry feels queasy'

Figure 2

In this system all pairs of *S*-worlds are *a*-related Call this a *maximally interrelated* system So in this system 'Harry is conscious' is true at S_{Dick} Indeed, at every *S*-world it is true that every functionally sentient creature is conscious There are no zombies in a maximally interrelated system

IV C Egocentric Presentism and the Logic of Points of View That is the *S*-world semantics for a logic of points of view It gives you, I hope, an intuitive sense of how the operators *from some point of view, from every point of view,* and so on, work As an egocentric presentist (pretend, for the moment, that I am one) I make use of these operators but believe the semantics to be misleading in one important respect I believe that all that exists is an *S*-world, S_{ME}, in which the experiences of one and only one person, the person I call 'me', are present That is it What makes the *S*-world semantics useful, though misleading, is that propositions containing the operators are true or false of this world as if it were part of a maximally-interrelated system of physically identical *S*-worlds, and truth conditions for such propositions were given by the rules above So it is true, for example, that *from Henry Kissinger's point of view* (Henry Kissinger's experiences are present), and that *from Henry Kissinger's point of view* (*from Condoleezza Rice's point of view* (Henry Kissinger's experiences are not present)), and that every functionally sentient creature is conscious

An egocentric presentist's attitude towards the *S*-world semantics, then, is very like a temporal presentist's attitude toward the standard four-dimensionalist semantics for tense logic The temporal presentist says that only the present moment exists, that that is all there is But for the purposes of understanding how tensed operators (operators like *it was once the case that* and *it will always be the case that*) work it may be useful to imagine that other moments exist, because propositions containing such operators are true or false of the present moment, as if it were part of a system of moments ordered by an access relation (the *earlier than* relation) and truth conditions for propositions containing the operators were given by the rules of the four-dimensionalist semantics for tense logic

It is also very like a modal fictionalist's attitude to the possible worlds semantics for modal logic [21] The modal fictionalist believes that propositions containing modal operators (like *it is possible that* and *it is necessary that*) are true or false of the actual world as if it were part of a wider system of possible worlds ordered by an access relation, and truth conditions for the operators were given by the possible

[21] See Gideon Rosen, "Modal Fictionalism," *Mind*, XCIX (1990) 327–54, and onwards

worlds semantics But, says the fictionalist, we should not let this fool us into thinking that a wider system really exists Only the actual world exists

IV D Egocentric Presentism and Egocentric-Hedonistic Considerations If egocentric presentism is right then all that there is is an *S*-world *S*-worlds are maximal states of affairs So, as an egocentric presentist, I can make peace between considerations of the greater good and egocentric-hedonistic considerations by taking *S*-worlds in which I (short-hand for 'the one with present experiences', remember) suffer to be worse simpliciter than *S*-worlds in which I do not, and *S*-worlds in which I prosper to be better simpliciter than *S*-worlds in which I do not

The Problem of Irreducibly Egocentric Preferences now disappears When I have such a preference, in the *After the Train Crash* case, for example, the two possibilities that seem open to me really do represent different ways for everything to be My desire that I not suffer amounts to a preference for the maximal state of affairs in which there is no present suffering, a simply better maximal state of affairs

What about the Generalization Problem? As an egocentric presentist I can make sense of the idea that other people have reason to be self-interested, even when their self-interest conflicts with my own Recall the *Competing for a Scarce Resource* case, for example In this case there seemed to be two ways for things to go

Possibility (1)	CJH gets the resource and there is present happiness	Jane misses out and there is absent misery
Possibility (2)	CJH misses out and there is present misery	Jane gets the resource and there is absent happiness

Since the state of affairs represented by possibility (1) is intrinsically better than the state of affairs represented by possibility (2), considerations of the greater good support *my* favoring it, but how do such considerations bear on Jane? I might think about this question in two different ways

On the one hand, I might think that considerations of the greater good support Jane favoring me getting the resource in just the way that they support me favoring me getting the resource If Jane takes the resource for herself then she brings about an intrinsically worse state of affairs, and that counts against her doing so

But there is another, better way to think about it When I treat another person as a practical deliberator, and ask whether such-and-such considerations support her doing so-and-so, I take her to be in

a particular deliberative context, *facing* choices, options, and alternatives This involves taking the deliberative context to be present But since Jane's experiences are not, have never been, and will never be present, she is not, has never been, and will never be a practical deliberator in this sense, so I cannot think directly about what considerations count I must instead think indirectly, by, so-to-speak, thinking within the scope of the *from Jane's point of view* operator—by thinking about what *from Jane's point of view* (is the case) When I do this I see that *from Jane's point of view* (if Jane gets the resource there will be present happiness and absent misery, if Jane misses out there will be present misery and absent happiness), so *from Jane's point of view* (it is intrinsically better that Jane get the resource), so *from Jane's point of view* (considerations of the greater good support Jane favoring the scenario in which she gets the resource) When I think about Jane as a practical deliberator, I see that considerations of the greater good support her looking out for herself

When a monadic tenser seeks to generalize time-bias, he will feel a similar kind of ambivalence Imagine that, in the *Bad and Worse Operations* case, I decide to act on Monday so as to ensure that I will have the nasty operation on Thursday, rather than the exquisitely agonizing one on Tuesday It is now Wednesday evening, and, being biased toward the future, I am already regretting my decision, wishing that I had brought my operation forward to Tuesday What should I now think about whether considerations of the greater good supported Monday's decision?

On one hand I might think that they weighed against my decision As things are there will be nasty suffering tomorrow If I had acted differently all suffering would be in the past So, in deferring the operation, I favored an intrinsically worse state of affairs over an intrinsically better one

But that is obviously the wrong way to think about it Thinking in the right way about the practical considerations that bore upon Monday's decision involves thinking about a deliberative context in which I face Monday's options and alternatives, which involves thinking about that deliberative context as present But Monday's deliberative context is not present For me to consider it so, I must think, so-to-speak, within the scope of the *it was the case that* operator I must think about how *it was the case that* (things are) But since *it was the case that* (if I choose to defer the operation there will be less suffering in the future), *it was the case that* (considerations of the greater good support my deferring the operation) When I think about the decision in the appropriate way, I see that it was supported by considerations of the greater good

Lastly, what about the Grounding Problem? Would the metaphysics support mild egocentric-hedonism by providing grounds for thinking that, all other things being equal, it is worse that suffering be mine than someone else's? I think so If you do not see this immediately, I urge you (really!) to perform this experiment

Trial By Kettle

> Today, many hundreds, if not thousands, of Russians will spill boiling water on their hands Pour boiling water on your own hand and compare your present discomfort with the absent discomfort of the northern-most Russian spiller Which is worse?

Your immediate reaction

> "This pain is dreadful, worse than any of those Russian pains "

may be tempered by a sober, reflective thought

> "My pain appears worse to me because I am more intimately acquainted with it It is present to me in a way that the northern-most Russian spiller's pain is not But he is more intimately acquainted with his pain It is present to him in a way that mine is not Since our situations are really symmetrical, I find, on reflection, that I have no grounds for thinking that my pain is worse simpliciter than his pain "

Well and good But this humbling thought is not available to an egocentric presentist For an egocentric presentist, the situations are not symmetrical It is not that my pain is present to me and his present to him Mine is present and his is absent That is part of the way things are So there is no reason to qualify or reassess my initial judgment

The thought experiment does not, of course, commit an egocentric presentist to extreme egocentric-hedonism—the view that only present suffering has any significance at all It identifies one factor that makes pleasure better and suffering worse This is quite compatible with thinking that, in evaluating the significance of suffering, there are many other relevant factors to consider—factors such as the *intensity* of the suffering, the *duration* of the suffering, and the *number* of sufferers It may be very tricky to give these factors precise weight Is it better that there be one hour of hand-scalding or four hours of thumb-scalding? Is it better that there be one scalded hand or four scalded thumbs?[22]

[22] The trickiness runs very deep Indeed, it may be that there is no satisfactory way of answering all such questions without appealing to an intransitive "better than" relation See Larry Temkin, "A Continuum Argument for Intransitivity," *Philosophy and Public Affairs,* XXV, 3 (1996) 175–210, and Alistair Norcross, "Comparing Harms Headaches and Human Lives," *Philosophy and Public Affairs,* XXVI, 2 (1997) 135–67

Is it better that there be absent suffering from hand-scalding or present suffering from thumb-scalding? But some cases are less tricky It is better that there be four hangnails than one crushed leg It is better that there be present suffering from a hangnail than absent suffering of leg-crushing After all, when there is an absent suffering from leg-crushing, and the victim is not a zombie, it is the case that *from someone else's point of view* (there is excruciating pain) And an egocentric presentist is free to take it that this matters, in just the way that a monadic tenser is free to take it that it matters that *it will be the case that* (there is excruciating pain) That is why empathy is instructive For an egocentric presentist, empathizing with an unfortunate involves imagining that the unfortunate has present experiences This involves viscerally imagining what *from the unfortunate's point of view* (is the case) And an egocentric presentist will care about the results of this exercise because he cares about what *from another person's point of view* (is the case)

So if I, a mild egocentric-hedonist, adopt egocentric presentism, in some cases (for example, comparing a state of affairs in which I suffer a hangnail with one in which someone else suffers a crushed leg) there will remain some conflict between egocentric-hedonistic considerations and considerations of the greater good But the residual conflict will not trouble me As an egocentric presentist, I do not take egocentric-hedonistic considerations to be considerations of a special, distinctive kind, that count towards my favoring one state of affairs over another even when I recognize that the other is better simpliciter than the one Once I have judged that it is better that there be a present hangnail than absent leg-crushing, it is not as if there is some further consideration—"but the hangnail will be mine!"—that has independent force I have already accounted for the presence of the hangnail, the fact that it is mine, in my judgment about which state of affairs is better simpliciter So when egocentric-hedonistic considerations and considerations of the greater good conflict, the former just fall away

Furthermore, when the trade-off between my interests and the interests of other people is more evenly balanced, as an egocentric presentist I can indulge my mildly egocentric preferences and attend to my own pleasures and pains while enjoying the psychological harmony of Louis XIV—the serene confidence that comes with believing that in doing what is good for me I am doing what is good full stop

CASPAR HARE

Massachusetts Institute of Technology

NOTES AND NEWS

The editors are saddened to note the death of Richard Rorty, on June 8, 2007, at the age of 75 Born in New York City in 1931, Rorty enrolled at Hutchins College at the University of Chicago at the age of 15 After serving in the United States Army and working on early computer systems, Rorty received his professional training from Yale University, completing his dissertation in 1956 Rorty's first faculty position came at Wellesley University in 1958 Rorty went on to positions as Professor of Philosophy at Princeton University, University Professor of the Humanities at the University of Virginia, and finally as Professor of Comparative Literature at Stanford University

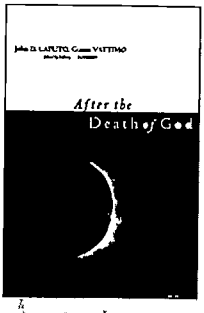
Rorty's books include *Philosophy and the Mirror of Nature* (1979), *Consequences of Pragmatism* (1982), *Contingency, Irony, and Solidarity* (1989), *Achieving Our Country Leftist Thought in Twentieth Century America* (1998), and four volumes of philosophical papers (1991, 1998, and 2007) Rorty received numerous awards and honors throughout his career, including honorary doctorates from eight universities in five countries, and held many visiting fellowships Rorty served as the President of the Eastern Division of the American Philosophical Association in 1979 and was an active member of the American Academy of Arts and Sciences, and the American Philosophical Society

After an early career of publishing erudite and well-respected essays, mostly in the Philosophy of Mind, Richard Rorty upturned the professional philosophical world in English speaking countries with his book *Philosophy and the Mirror of Nature* in which he presented his own historical account of Epistemology as governed by a notion of "objectivity," which he wished to repudiate, and in subsequent essays, he argued equally against a view in the Philosophy of Language which gave a central place to the notion of "truth " Rorty provoked philosophers by claiming that in the last few decades their intellectual "subject" had, to its detriment, transformed itself into a "profession," with coteries advancing their careers by narrowing its themes, tyrannically limiting its rhetoric and idiom and style of argument, and referring and responding only to each other's work in the manner of a small sect Despite a hostile and defensive reaction among some philosophers in the immediate aftermath of its publication, *Philosophy and the Mirror of Nature* is widely acknowledged among many who now concede that Philosophy needed a jolt of that kind, and some of his most interesting contemporaries such as Donald Davidson, John McDowell, and Bernard Williams themselves responded with seriousness and a sense of urgent debate to matters of substance raised by Rorty The book is now viewed as an original and substantial work, not only among philosophers in other traditions, such as on those on the European continent who felt in some way vindicated by it in their opposition to so-called "analytical" philosophy, but by many analytic philosophers as well

Apart from his contributions to these subjects, Rorty brought to the philosophical study of politics and culture a very distinctive point of view, described by him sometimes in deliberately conventional terms as "liberal" and "Western," but this vein of his work was filled with ardor, wit, shrewdly executed rhetorical skill, and the desire to usher liberal politics back to its "Old Left" margins His political ideas were not discontinuous with his epistemological claims It was a mark of his intellectual ambition that he always tried to integrate his diverse themes, even if it often meant that he would strain to make his heroes—Wittgenstein, Heidegger, Davidson, Kuhn, Dewey—all come out saying roughly the same thing An abiding mark of his influence among philosophers is our growing recognition that—done at their best and their most interesting—European Philosophy and "Analytical" Philosophy really pursue the same subjects and interests, even if they do so in dramatically different idioms The death of a philosopher so willing to swim against the stream is a great loss to the discipline

COLUMBIA, *Refer to us*

Read book excerpts at www columbia edu/cu/cup



After the Death of God

John D Caputo & Gianni Vattimo

Edited by Jeffrey W Robbins

978 0 231 14124 6 - cloth - $24 50

Insurrections Series

On Suicide Bombing

Talal Asad

978 0 231 14152 9 cloth $19 95

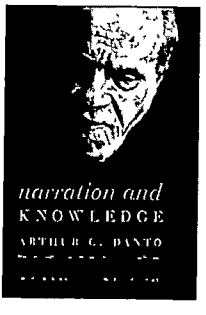
The Wellek Library Lectures

Confucian Tradition and Global Education

Wm Theodore de Bary

With contributions by Cheung Chan Fai and Kwan Tze wan

978 0 231 14120 8- cloth $32 50



Narration and Knowledge

Arthur Danto

With a new introduction by Lydia Goehr and a new conclusion by Frank Ankersmit

978-0 231-13822-2 cloth $69 50

978 0-231 13823 9 paper $24 50

Columbia Classics in Philosophy

Narrating Evil

A Postmetaphysical Theory of Reflective Judgment

Maria Pia Lara

978 0 231 14030 0 cloth $29 50

New Directions in Critical Theory

Neuroscience and Philosophy

Brain, Mind, and Language

Maxwell Bennett, Daniel Dennett, Peter Hacker, & John Searle

Introduction and conclusion by Daniel N Robinson

978 0 231 14044-7 - cloth - $24 50

Being Human

Historical Knowledge and the Creation of Human Nature

Roger Smith

978 0 231 14166 6 - cloth $29 50

The Universal (In the Realm of the Sensible)

Beyond Continental Philosophy

Dorothea Olkowski

Introduction and conclusion by Daniel N Robinson

978 0 231 14198 7 cloth - $37 50

NOW IN PAPER

The Future of Religion

Revised Edition

Richard Rorty & Gianni Vattimo

Edited by Santiago Zabala

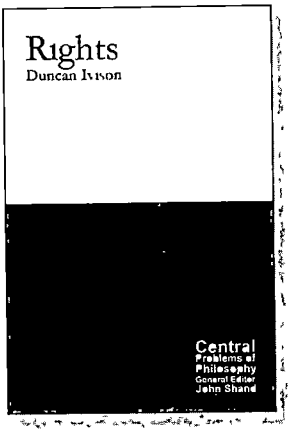
978 0 231-13495-8 - paper -$17 95

Glimmer of a New Leviathan

Total War in the Realism of Niebuhr, Morgenthau, and Waltz

Campbell Craig

978 0 231-12349 5 - paper -$24 50

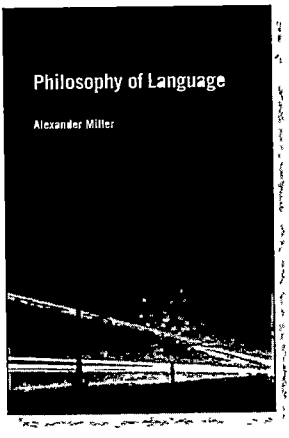


Rights

Duncan Ivison

978-0-7735 3329 5 $22 95 paper | 978-0-7735-3328-8 $80 cloth

An accessible introduction to the history and philosophy of rights

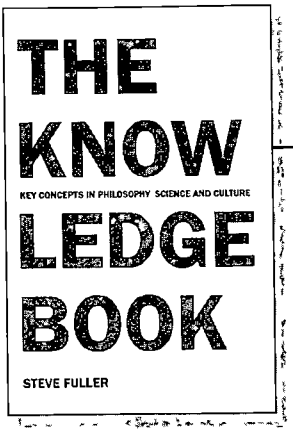


Philosophy of Language
Second Edition

Alex Miller

978-0-7735-3338-7 $22 95 paper

This second edition includes new material on Chomsky, Wittgenstein, and Davidson as well as new chapters on the causal theory of reference, possible worlds semantics, and semantic externalism



The Knowledge Book

Key Concepts in Philosophy, Science, and Culture

Steve Fuller

978 0 7735-3347-9 $22 95 paper | 978-0-7735-3346 2 $80 cloth

"An invaluable resource for anyone interested in exploring the social life of knowledge Fuller's unrivalled concision, integrated with the book's unique structure, invites readers to imagine, and to chart, the future of knowledge"

–James H Collier, Virginia Tech

MQUP McGILL-QUEEN'S UNIVERSITY PRESS
www mqup ca

New from OXFORD

New in Paperback!

On Desire

Why We Want
What We Want

WILLIAM B IRVINE

"A program of illumination whose goal is to help the reader 'master desire.' That human life depends on desire, or is at the least inseparable from it, Irvine does not dispute, but his purpose is not to glorify this essence that artists celebrate. Exposes 'the secret life of desire' how we experience our wants, from those for basic animal requirements, like food, to abstract goals—the esteem of our fellows, for example—that may feel as necessary as food to complicated creatures who cannot live by bread alone." —*The New York Times Book Review*

2007 336 pp paper $13.95

Morality and Self-Interest

Edited by PAUL BLOOMFIELD

The relationship between morality and self-interest is a perennial one in philosophy. For Plato, Hobbes, Kant, Aristotle, Hume, Machiavelli, and Nietzsche, it lay at the heart of moral theory. But little of the contemporary work has been published in book format. Bloomfield's edited volume will be the first such book devoted to morality and self-interest, presenting new, commissioned articles on this subject by some of the top philosophers working today.

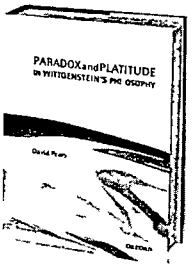
2007 368 pp paper $24.95 cloth $74.00

New Pragmatists

Edited by CHERYL MISAK

The classical pragmatists, Charles Sanders Peirce, William James, and John Dewey, put forward views of truth, rationality, and morality that they took to be connected to, and good for, our practices of inquiry and deliberation. In this volume, some of our very best contemporary philosophers explore this and develop the pragmatist project, showing that pragmatism is a strong current in philosophy today.

2007 216 pp $45.00

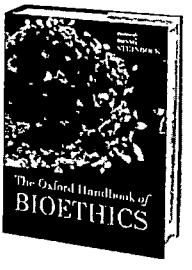


Paradox and Platitude in Wittgenstein's Philosophy

DAVID PEARS

This is a concise and readable study of five intertwined themes at the heart of Wittgenstein's thought, written by one of his most eminent interpreters. David Pears offers penetrating investigations and lucid explications of some of the most influential and yet puzzling writings of twentieth-century philosophy. He focuses on the idea of language as a picture of the world, the phenomenon of linguistic regularity, the famous "private language argument", logical necessity, and ego and the self.

2007 150 pp $35.00



The Oxford Handbook of Bioethics

Edited by
BONNIE STEINBOCK

Bonnie Steinbock presents the authoritative, state-of-the-art guide to current issues in bioethics, covering 30 topics in original essays by some of the world's leading figures in the field, as well as by some newer 'up-and-comers'. Anyone who wants to know how the central debates in bioethics have developed in recent years, and where the debates are going, will want to consult this book.

(Oxford Handbooks)

2007 768 pp $150.00

OXFORD
UNIVERSITY PRESS

Prices are subject to change and apply only in the US. To order, please call 1-800-451-7556. In Canada, call 1-800-387-8020. Visit our website at www.oup.com/us

New from Stanford University Press

Psyche
Inventions of the Other, Volume I

JACQUES DERRIDA, Edited by PEGGY KAMUF and ELIZABETH G ROTTENBERG

"This monumental collection of essays shows Derrida at his brilliant best, across a vast and diverse range of topics, texts, authors and manners *Psyche* is among the richest and most diverse of all Derrida's books, and a testimony to the extraordinary depth and vigor of deconstructive thought"

—Geoffrey Bennington, Emory University

Meridian Crossing Aesthetics
$24 95 paper $65 00 cloth

The Equivocation of Reason
Kleist Reading Kant

JAMES PHILLIPS

Kleist is a famous misreader of Kant, but this study pitches the latter's principles against the more restricted scope of his own examples in order to develop an ethics and an account of the sublime in keeping with Kleist's literary works
$50 00 cloth

Stanford University Press

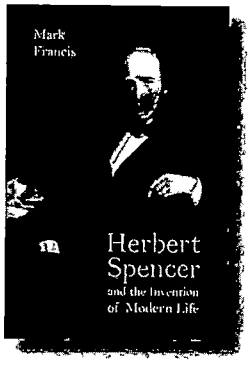
800.621.2736
www sup.org

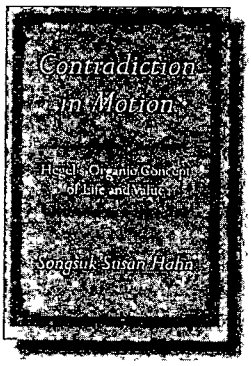
You can now subscribe to the ***Journal of Philosophy*** using

PayPal

To learn more about this easy new option, click on the PayPal icon at **www.jphil.org.**

NEW TITLES IN PHILOSOPHY FROM CORNELL UNIVERSITY PRESS





HERBERT SPENCER AND THE INVENTION OF MODERN LIFE
MARK FRANCIS

"A major new study of Spencer, revealing aspects of his personality and thought previously little explored. It is an impressive work of scholarship and interpretation that no scholar of nineteenth-century thought can afford to neglect."

—David Boucher, Cardiff University
464 pages, $45 00 cloth

CONTRADICTION IN MOTION
Hegel's Organic Concept of Life and Value
SONGSUK SUSAN HAHN

"Hahn has her finger on the pulse of Hegel scholarship, and this is a timely book, as well as one taking on an old and difficult problem in some refreshing ways"

—Allen Wood, Stanford University
240 pages, $49 95 cloth

THE WISDOM TO DOUBT
A Justification of Religious Skepticism
J. L. SCHELLENBERG

"This is a brilliant work, full of original thinking and unhurriedly persuasive argument The project it carries forward is a major new departure in the philosophy of religion, and should break much of the deadlock in contemporary debate."

—Terence Penelhum, University of Calgary
344 pages, $49 95 cloth

RACE OR ETHNICITY?
On Black and Latino Identity
EDITED BY JORGE J. E. GRACIA

"This book is interesting, timely, and worthwhile because it addresses contemporary issues regarding highly contestable concepts that are usually taken for granted in ordinary as well as in professional conversations about race and ethnicity"

—Vicente Medina, Seton Hall University
304 pages, $22 95 paper

New in Paperback!
THE ROOTS OF EVIL
JOHN KEKES

"This is an interesting, systematic, nondogmatic, and informed attempt to make sense of evil on secular grounds"

—*Times Literary Supplement*
278 pages, $19 95 paper

www.cornellpress.cornell.edu

CAMBRIDGE

What's New in Philosophy?

Alvin Plantinga
Edited by Deane-Peter Baker
Contemporary Philosophy in Focus
$75.00 Hb. 978-0-521-85531-0 246 pp
$24.99 Pb. 978-0-521-67143-9

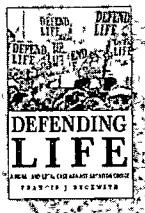
Defending Life
A Moral and Legal Case Against Abortion Choice
Francis J. Beckwith
$75.00 Hb 978-0-521-87084-9. 312 pp
$22.99 Pb. 978-0-521-69135-2

Identities, Affiliations, and Allegiances
Edited by Seyla Benhabib, Ian Shapiro, and Danilo Petranovich
$85.00 Hb 978-0-521-86719-1 434 pp
$29.99 Pb. 978-0-521-68693-8

The Cambridge Companion to Plato's *Republic*
Edited by G. R. F. Ferrari
Cambridge Companions to Philosophy
$80.00 Hb 978-0-521-83963-1 560 pp
$29.99 Pb 978-0-521-54842-7

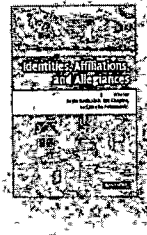
Ethics and Business
An Introduction
Kevin Gibson
Cambridge Applied Ethics
$85.00 Hb. 978-0-521-86379-7. 274 pp
$29.99 Pb. 978-0-521-68245-9

Objectivity and the Rule of Law
Matthew Kramer
Cambridge Introductions to Philosophy and Law
$75.00 Hb 978-0-521-85416-0 262 pp
$27.99 Pb 978-0-521-67010-4

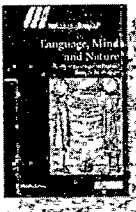


Wittgenstein's Apprenticeship with Russell
Gregory Landini
$96.00 Hb. 978-0-521-87023-8 312 pp.

The Young Karl Marx
German Philosophy, Modern Politics, and Human Flourishing
David Leopold
Ideas in Context
$105.00 Hb. 978-0-521-87477-9 338 pp

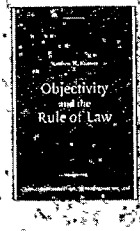


Language, Mind and Nature
Artificial Languages in England from Bacon to Locke
Rhodri Lewis
Ideas in Context
$96.00 Hb 978-0-521-87475-5 286 pp.



Hegel's Critique of Metaphysics
Beatrice Longuenesse
Translated by Nicole J. Simek
Modern European Philosophy
$85.00 Hb. 978-0-521-84466-6. 268 pp.

Morality in a Technological World
Knowledge as Duty
Lorenzo Magnani
$85.00 Hb 978-0-521-87769-5. 304 pp



Unmodern Men in the Modern World
Radical Islam, Terrorism, and the War on Modernity
Michael J. Mazarr
$75.00 Hb 978-0-521-88175-3 312 pp
$19.99 Pb 978-0-521-71291-0

Prices subject to change

www.cambridge.org/us/philosophy

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE

What's New in Philosophy?

Weber, Habermas and Transformations of the European State
Constitutional, Social, and Supra-National Democracy
John P McCormick
$96 00 Hb 978-0 521 81140 8 318 pp

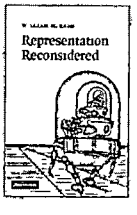
The Globalization of Ethics
Religious and Secular Perspectives
Edited by William Sullivan and Will Kymlicka
$80 00 Hb 978 0-521-87335-2 320 pp
$27 99 Pb 978 0 521-70021 4

Proclus Commentary on Plato's *Timaeus*
Volume 1 - Book I Proclus on the Socratic State and Atlantis
Edited and translated by Harold Tarrant
$128 00 Hb 978 0-521-84659 2 358 pp
Volume 3 - Book 3
Edited and translated by Dirk Baltzly
Proclus Commentary on Plato's Timaeus
$91 00 Hb 978 0 521 84595 3 218 pp

Kant's *Groundwork of the Metaphysics of Morals*
A Commentary
Jens Timmermann
$91 00 Hb 978 0 521 86282 0 224 pp

Making Prehistory
Historical Science and the Scientific Realism Debate
Derek Turner
Cambridge Studies in Philosophy and Biology
$91 00 Hb 978 0 521 87520-2 238 pp

Representation Reconsidered
William M Ramsey
$91 00 Hb 978-0 521 85987-5 268 pp



Virtue Ethics and Consequentialism in Early Chinese Philosophy
Bryan Van Norden
$90 00 Hb 978 0 521 86735 1 426 pp

Is the Welfare State Justified?
Daniel Shapiro
$80 00 Hb 978-0-521 86065-9 344 pp
$27 99 Pb 978 0-521 67793 6

Dependence Logic
A New Approach to Independence Friendly Logic
Jouko Väananen
London Mathematical Society Student Texts
$125 00 Hb 978 0-521 87659 9 236 pp
$45 00 Pb 978 0 521 70015-3

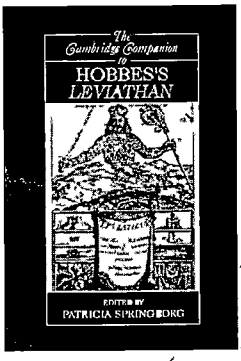
The Cambridge Companion to Hobbes's *Leviathan*
Edited by Patricia Springborg
Cambridge Companions to Philosophy
$85 00 Hb 978 0-521 83667-8 560 pp
$29 99 Pb 978 0-521 54521 1

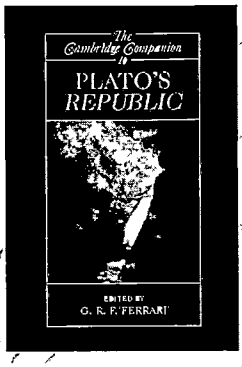
Prices subject to change

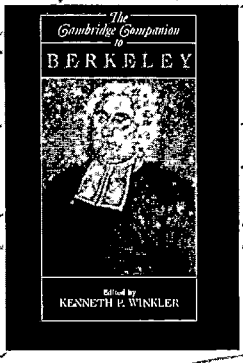
www cambridge org/us/philosophy

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE







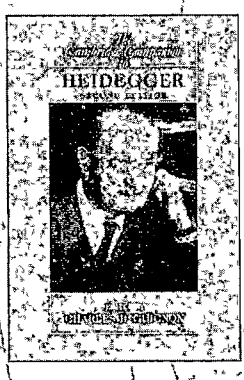
Cambridge Companions Online

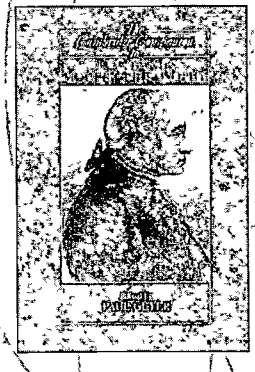
The acclaimed series is now available electronically!

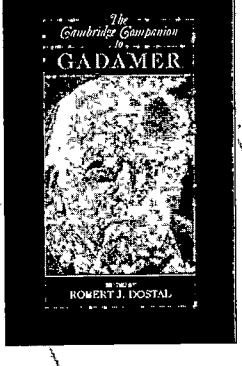
Integrating over 250 titles from the renowned print series into a dynamic, fully-searchable interface, **Cambridge Companions Online** offers unbeatable functionality as both a teaching and research tool

The online collection is available in three packages

- The **Philosophy, Religion** and **Culture** Companions Collection
- The **Literature** and **Classics** Companions Collection
- The **Complete Companions Collection** (covering literature, classics, philosophy, religion, and cultural studies)







For a FREE institutional trial, write to ccol@cambridge.org

www.cambridge.org/us

CAMBRIDGE UNIVERSITY PRESS

THE JOURNAL OF PHILOSOPHY

VOLUME CIV, NUMBER 8
AUGUST 2007

page

375 *Quining Naturalism* — Huw Price

403 *Safety, Content, Apriority, Self-Knowledge* — David Manley

BOOK REVIEWS

424 Peter Unger *All the Power in the World* — Stephen Mumford

432 NEW BOOKS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Wolfgang Mann, Mary Mothersill, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C Varzi *Consulting Editors* David Albert, James T Higginbotham, Robert May, Charles D Parsons, Wilfried Sieg *Managing Editor* John Smyhe

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President*, Arthur C Danto, *Vice President*, Akeel Bilgrami, *Secretary*, Michael J Mooney, *Treasurer*, Liaquat Ahamed, *Other Trustees* Lee Bollinger, Leigh S Cauman, Barbara Gimbel, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro

All communications to the Editors and Trustees and all manuscripts may be sent to John Smyhe, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

You may also visit our website at www journalofphilosophy org

THE JOURNAL OF PHILOSOPHY

2007

SUBSCRIPTIONS (12 issues)

Individuals	$45 00
Libraries and Institutions	$100 00
Libraries and Institutions (print and online)	$180 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for advertising rates, ad space is limited, so ad reservations are required

Published monthly as of January 1977, typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027 (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U S copyright law All rights reserved For more information about our copyright policies, please consult our January issue or contact us

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc , MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027 Periodicals postage paid at New York, NY, and other mailing offices

POSTMASTER. Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027

© Copyright 2007 by the Journal of Philosophy, Inc

ISSN 0022–362X

THE JOURNAL OF PHILOSOPHY
VOLUME CIV, NO 8, AUGUST 2007

UNIVERSITY OF CALCUTTA
CAMPUS

QUINING NATURALISM*

Scientific naturalism is a metaphysical doctrine, a view about what there is, or what we ought to believe that there is It maintains that natural science should be our guide in matters metaphysical the ontology we should accept is the ontology that turns out to be required by science W V Quine is often regarded as the doyen of scientific naturalists, though the supporting cast includes such giants as David Lewis and J J C Smart

An alternative view offers a causal criterion for ontological commitment we should be realists about whatever manifests itself in virtue of having effects In particular, perhaps, we should believe in the existence of whatever figures in good causal explanations of observed phenomena, and of our experiences and beliefs in general Thus electrons play a role in good causal explanations of our beliefs "about electrons," presumably, and so we should be realists about electrons, but values do not seem to figure in causal explanations of our evaluative beliefs, so we should not be realists about values

These two criteria for realism are not the same, of course The first is often thought to provide a reason for doubting that the second is a necessary condition, on the grounds that entities that are not causally efficacious—numbers, perhaps, or possible worlds—may be indispensable to science There is much agreement, however Much of what figures in explanations of our beliefs also figures in natural science All of it does so, as long as explanations of our beliefs are *ipso facto* scientific explanations In that case, proponents of second criterion

*I am grateful to David Macarthur, Steven Savitt, and Amie Thomasson, and to audiences in Rome and Canberra, for much helpful criticism of earlier versions of this paper I am also grateful to the Australian Research Council, for research support

0022-362X/07/0408/375–402 © 2007 The Journal of Philosophy, Inc

are entitled to regard themselves as strictly more stringent naturalists than their permissive Quinean cousins

David Armstrong calls the causal criterion the Eleatic Principle,[1] and takes it, *contra* Lewis, to favor fictionalism about possible worlds Other notable advocates of the Eleatic criterion include Hartry Field, who takes it to provide support for fictionalism about mathematics, and Simon Blackburn, who suggests that it marks the line at which realism should give way to quasi-realism [2]

In this paper, I want to criticize both criteria for realism, and both resulting forms of naturalism—"Quinean naturalism" and "Eleactic naturalism," as I will call them Neither provides an appropriate normative criterion for ontological commitment, in my view, and domains which fail both tests may be on a par with domains which pass both Thus my project is make a case for the removal, or at least for a radical relocation, of the boundary that both forms of naturalism draw between real and unreal, respectable and unrespectable, in the ontological realm And I want to make the case, by and large, from Quinean materials

I PALE METAPHYSICS

Think of the distinction between realism and irrealism as like that between illumination and shadow, in a black and white image The Quinean and Eleatic versions of naturalism offer us with pictures in which the light falls bright on the entities required by natural science, or by causal explanations, but all else lies in shadow I want to defend an alternative image, in which neither region is highlighted in this distinctive way

In principle, there are three ways to adjust the naturalists' image, to produce such a result We might increase the brightness overall, revealing structure in parts of the image previously dark We might reduce the brightness overall, losing the structure in regions previously light Or we might simply reduce the contrast overall, producing a paler and more nuanced image, in shades of grey The last manipulation is the one I favor [3]

[1] Armstrong, *A World of States of Affairs* (New York Cambridge, 1997), p 41 See also Graham Oddie, "Armstrong on the Eleatic Principle and Abstract Entities," *Philosophical Studies,* XLI (1982) 285–95

[2] See Field, *Science without Numbers A Defence of Nominalism* (Princeton University Press, 1980), Blackburn, "Truth, Realism, and the Regulation of Theory," in *Essays in Quasi-Realism* (New York Oxford, 1993), pp 15–34

[3] Or strictly, it is the first of two manipulations I want to recommend The second adds some new hues, while keeping the contrast low

The metaphor is self-consciously Quinean, of course Recall Quine's famous trope for the rejection of a distinction between factual and conventional truths

> The lore of our fathers is a pale gray lore, black with fact and white with convention But I have found no substantial reasons for concluding that there are any quite black threads in it, or any white ones [4]

I have borrowed Quine's metaphor because I, too, want to call into question a well-entrenched distinction—and to do it, by and large, by Quinean methods My target is the high-contrast metaphysical picture offered to us by naturalism In other words, in the sense of Daniel Dennett's *The Philosophical Lexicon,* and hence of his "Quining Qualia,"[5] I want to "quine" the distinction that naturalists draw between light and dark, real and nonreal, in the ontological realm But my target is Quine's own scientific naturalism, as well as its Eleatic cousin So there is irony, as well as homage, in my metaphorical mimicry

My view differs from these two versions of naturalism in two ways first, in "reducing the contrast," in deflating the conception of what is at stake in these metaphysical matters, and second, in maintaining that in so far as there is any distinction worth making, it does not fall where (either version of) naturalism takes it to fall In the first matter, I am going to argue that I am entitled to treat Quine as an ally One of my main tasks will be to distinguish two conceptions of Quine's prescription for metaphysics and ontology, which seem to me to have been systematically confused Only the weaker conception is really defensible by Quinean lights, I shall argue, and this amounts to the deflationary view

In the second matter, Quine is not officially an ally, but ought to be, in my view Even if the stronger conception of what is at stake were sustainable—if ontology were a matter of black and white, so to speak—there would be little in Quine to support the naturalists' account of where the line should be drawn

My view is close to that of Rudolf Carnap's "Empiricism, Semantics, and Ontology "[6] Carnap defends both a pale, deflationary conception of ontology, in my sense, and also a kind of pluralism that challenges

[4] Quine, "Carnap and Logical Truth," in *Ways of Paradox and Other Essays,* 2nd edition (Cambridge Harvard, 1976), pp 107–32, at p 132

[5] Dennett, "Quining Qualia," in A Marcel and E Bisiach, eds , *Consciousness in Modern Science* (New York Oxford, 1988), pp 42–77

[6] Carnap, "Empiricism, Semantics, and Ontology," *Revue Internationale de Philosophie,* IV (1950) 20–40, reprinted in *Meaning and Necessity A Study in Semantics and Modal Logic,* 2nd enlarged edition (Chicago University Press, 1956), pp 205–21 (Page references here are to the latter version)

the naturalist's privileged conception of the place of science In my terms, in other words, he is a champion of pale but multi-colored metaphysics It may seem foolhardy to pit such a champion against Quinean naturalism, of all targets Didn't Quine himself demolish Carnap's view—using, what is more, the very weapon whose metaphors I have mimiced, namely, the rejection of a distinction between matters of fact and matters of linguistic convention? However, like the widespread belief that Quine made the world safe for less pale forms of metaphysics, this assessment of what he achieved against Carnap is mistaken, in my view

As I said, I want to argue that Quine's own position on ontological commitment, and the relation of philosophy to science, in fact provides little support for "Quinean" naturalism On the contrary, it favors, or at least leaves open, a view much closer to that of Carnap I will explain why this view is untouched by Quine's criticisms of Carnap, and argue that the most significant disagreement between Quine and Carnap concerns the plurality of the functions of language, and hence of ontological commitment Carnap is at least implicitly committed to pluralism, and Quine opposed to it When the issues in question are properly understood, however, Quine's objections to Carnap on this matter turn out to miss the target that matters Quine's monism then seems not only unsupported, but in tension, at least *prima facie,* with his naturalistic conception of the project of philosophy

Exploring these issues will lead us to the suggestion that the Eleactic criterion might provide a way of keeping pluralism under control—of distinguishing a privileged domain of ontological commitment, which alone deserves our full allegiance I will conclude by arguing that this suggestion, too, turns out to be untenable, by the lights of the broadly Quinean considerations I will be invoking against Quinean naturalism itself Thus the problems of the Quinean criterion for naturalism are not an argument in favor of the stricter Eleatic criterion On the contrary, they are an argument in favor of Carnapian tolerance

II A WIDER SHADE OF PALE?

Carnap thought that much of traditional metaphysics and ontology rests on a mistake In explaining why, he relies on the notion of a linguistic framework Roughly, a linguistic framework is the set of rules (supposedly) governing the use of a group of terms and predicates—say, the terms we use in talking about medium-sized objects, or in talking about numbers Carnap thought that adopting such a framework, or way of talking, typically brings with it ontological methods and questions These are "internal" questions, questions that arise within the

framework, and their nature depends on the framework in question They may be empirical, as in science, or logical, as in mathematics

However, Carnap continues, these internal questions do not include the metaphysical questions typically asked by philosophers "Are there material objects?," for example, or "Are there numbers?" Carnap says that in this form these "external" questions are simply mistakes they cannot be asked because they are "framed in a wrong way" (*op cit*, p 207) The only legitimate external questions are pragmatic in nature Should we adopt this framework? Would it be useful?[7]

Carnap thus becomes a pluralist about ontological commitment—explicitly so, in the sense that he associates distinct ontological commitment with distinct linguistic frameworks, and at least implicitly so in a deeper "functional" or pragmatic sense After all, the key to Carnap's accommodation of abstract entities is the idea that the framework that introduces talk of such entities may serve different pragmatic purposes from the framework that introduces talk of physical objects—and this could only be so if there is some sense in which the two frameworks "do different jobs " A corollary of this functional pluralism is that it may turn out that neither natural science nor the project of causal explanation provide the only frameworks in which we have a pragmatic need for ontological claims In principle, then, Carnap provides grounds for rejecting both forms of naturalism

However, Carnap's view is not simply a recipe for more inclusive realism—a way of turning up the brightness overall, as I put it earlier For if what is meant by realism is a metaphysical view, in the old sense, then Carnap's position amounts to a *rejection* of all such views By that realist's lights, then, Carnap's view is a form of global irrealism Yet his view is not simply a recipe for turning down the brightness overall, either It is a third position, which rejects the high-contrast terms in which the image is typically presented Here is Carnap's own negotiation of this critical point, from "Empiricism, Semantics, and Ontology"

> The non-cognitive character of the questions which we have called here external questions was recognized and emphasized already by the Vienna Circle under the leadership of Moritz Schlick, the group from which the movement of logical empiricism originated Influenced by ideas of Ludwig Wittgenstein, the Circle rejected both the thesis of the reality of the external

[7] A useful way to put Carnap's point is to appeal to the use–mention distinction Legitimate *uses* of the terms 'number' and 'material object' are necessarily internal, for it is conformity (more-or-less) to the rules of the framework in question that constitutes use But as internal questions, as Carnap notes, these questions could not have the significance that traditional metaphysics takes them to have Metaphysics tries to locate them somewhere else, but thereby commits a use–mention fallacy The only legitimate external questions simply *mention* the terms in question

> world and the thesis of its irreality as pseudo-statements, the same was the case for both the thesis of the reality of universals (abstract entities, in our present terminology) and the nominalistic thesis that they are not real and that their alleged names are not names of anything but merely *flatus vocis* (It is obvious that the apparent negation of a pseudo-statement must also be a pseudo-statement) It is therefore not correct to classify the members of the Vienna Circle as nominalists, as is sometimes done However, if we look at the basic anti-metaphysical and pro-scientific attitude of most nominalists (and the same holds for many materialists and realists in the modern sense), disregarding their occasional pseudo-theoretical formulations, then it is, of course, true to say that the Vienna Circle was much closer to those philosophers than to their opponents (*op cit*, p 215) [8]

Thus Carnap is not only an ontological pluralist, but also a champion of pale metaphysics, in my sense

But an unsuccessful champion, in many eyes According to a popular version of the history of twentieth-century philosophy, Quine was the savior of a more robust metaphysics, slayer of positivist demons hell-bent on exsanguinating the entire subject With one hand (the story goes), Quine wrote "On What There Is," and thus gave Ontology a life-saving transfusion, with the other, he drove a stake through the heart of "Empiricism, Semantics, and Ontology," and thus dispatched the last incarnation of the Viennese menace

I am exaggerating, of course, but not much—here is Hilary Putnam's description of the first of these achievements

> "How come," the reader may wonder, "it is precisely in *analytic* philosophy—a kind of philosophy that, for many years, was *hostile* to the very word 'ontology'—that Ontology flourishes?"
>
> If we ask *when* Ontology became a respectable subject for an analytic philosopher to pursue, the mystery disappears It became respectable in 1948, when Quine published a famous paper titled "On What There Is " It was Quine who single handedly made Ontology a respectable subject [9]

I want to argue that this orthodox philosophical history is mistaken in two ways

(1) The metaphysics that Quine revived is itself a pale zombie, not the sanguine, beefy creature that positivists since Hume had being trying to put down [10]

(2) Quine's stake missed the heart of Carnap's metaphysics-destroying doctrine completely, merely lopping off some inessential append-

[8] Carnap is here endorsing the views he ascribes to the Vienna Circle, of course

[9] Putnam, *Ethics without Ontology* (Cambridge Harvard, 2004), pp 78–79

[10] This needs qualification In one sense, Quine actually revived the more beefy kind of metaphysics Certainly it has been behaving as if it were very much alive But I would

> ages, and leaving the creature, if anything, stronger than before The twin-chambered heart of Carnap's view comprises, first, a deflationary view of metaphysics, with which Quine concurs, and second, a pluralism about the functions of existentially quantified discourse, with which Quine does not concur, but against which he offers no significant argument

III HOW BEEFY IS QUINE'S RECIPE FOR ONTOLOGY?

To begin the case for the first proposition, let us ask what sort of a subject it is, this "Ontology" that Quine is supposed to have made respectable Here is a recent three-line summary by Stephen Yablo

> Quine takes existence questions dead seriously He even outlines a program for their resolution Look for the best overall theory—best by ordinary scientific standards or principled extensions thereof—and then consider what has to exist for the theory to be true [11]

Yablo can afford to be succinct Quine's program for ontology is as familiar as almost anything in contemporary analytic philosophy, and Yablo is entitled to assume that his readers know what he means

Yet the familiarity of Quine's program conceals a trap, in my view In reality, there are two very different ways to construe the program, which need to be distinguished a lot more carefully than has usually been the case (Quine himself is one of those at fault, I think) Roughly, there is a "thin," or modest, reading of the program and a "thick," or ambitious reading I want to distinguish the two, and to argue that only the former really stands up as an adequate interpretation of Quine—though many people are committed, at least implicitly, to the stronger reading

To forestall a possible objection, I want to emphasize that in one sense, the distinction between these two readings makes little difference to the conclusion for which I am aiming My targets are the criteria that Quinean and Eleatic naturalists offer for ontological commitment, and under both readings of his program, in my view, Quine fails to offers any good argument for thinking that the boundary between light and dark should be drawn where these naturalists want it drawn (More on this later)

Under the thin reading, however, it is also doubtful whether the boundary between light and dark could possibly have the significance

---

prefer to say that it is actually as dead as Carnap left it, but that many of its practitioners—encouraged in part by the misinterpretation of Quine I am about to describe—simply have not noticed

[11] Yablo, "Does Ontology Rest on a Mistake?" *Proceedings of the Aristotelian Society,* Supp Vol LXXII (1998) 229–61, at p 230

that naturalists take it to have Under the thin reading, Quinean ontological commitment is already pale grey commitment—metaphysics has already lost its regions of black and white On the thin reading, in other words, Quine's brand of ontological commitment differs little in tone from Carnap's—while if it differs in application, in being restricted to the commitments needed in science, it thus rests on foundations that seem hard to defend, in the modest as well as in the ambitious version of the program

The distinction between thin and thick readings is important in another way, too It seems to me that many who appeal to Quine in support of their metaphysical investigations rely on the thick, ambitious reading, while at the same time displaying a kind of false modesty—helping themselves to a cloak of plain-speaking ontological frugality that belongs to the thin reading Metaphysics thus gets away with working both sides of the street, because the two readings are not properly distinguished It is therefore worth taking the trouble to draw the distinction, and to show that only the thin reading can really be regarded as legitimate, by Quine's own lights

I am going to proceed by offering extreme versions of the thin and thick readings, in order to highlight two kinds of constraint on an adequate interpretation of Quine's program for ontology My intention is that as we chart an acceptable course between these extremes—an under-nourished Scylla and a beefy Charybdis, as it were—we will find that the resulting view is clearly modest rather than ambitious, in the relevant respects In other words, I will begin by drawing the distinction between modest and ambitious readings in a caricatured form, and establish that in this form, only the modest reading is acceptably Quinean In this form, however, the modest reading is unacceptably trivial There are ways of making it less trivial, but so long as we keep our eyes on the dangers of Charybdis—on what it is that made our initial thick reading unacceptably ambitious—it will be clear, I will argue, that permissible enhancements of the trivial reading remain modest in the relevant respects Quine's program cannot be a recipe for thick or ambitious metaphysics

In order to highlight the threat posed by the thin interpretation to the view that Quine saves ontology as a substantial discipline, let us begin by reformulating Yablo's summary of Quine's program, so that it becomes a program for resolving factual matters in general, rather than specifically matters of ontology

> Quine takes [factual] questions dead seriously He even outlines a program for their resolution Look for the best overall theory—best by ordinary scientific standards or principled extensions thereof—and then consider [how the facts have to be] for the theory to be true

Now the initial challenge Hasn't the program become trivial? If the best overall theory says that *P*, then this is how the facts have to be for the theory to be true It has to be a fact that *P* So here is what the program seems to amount to *Look at what the scientific experts say, and take the facts to be what they say that they are* Some program!

In its extreme version, then, the modest reading amounts to the view that the investigation of reality is nothing more or less than what the lead players do on the scientific stage If philosophers want to do ontology first-hand, they should become scientific experts—do what scientists do, and do it well Otherwise, the only way to do it is second-hand, as spectators, by watching those who do it first-hand—by basing one's views on the views of the scientific experts (Those who can, do science Those who cannot, read *Science*) Nobody will deny that this program ("Look at what the scientific experts say, and take the facts to be what they say that they are") is a modest program for metaphysics—self-effacing to the point of extinction, as an intellectual pursuit in its own right However, everybody will deny, quite rightly, that it is all that Quine offers to metaphysics There are at least three ways in which it may be held to be a nontrivial matter how the world has to be, for the scientist's claim that *P* be true

First, some philosophers will say that there is often, perhaps always, a nontrivial issue about what *makes* such a claim true (if it is true) The task of philosophy, then, is the search for *truth makers* for the best overall theories given to us by science Second, even philosophers wary of the notion of truth makers and truth-making may allow that there is often a nontrivial issue about the *interpretation* of scientific theories accepted as true Granted that quantum mechanics is true, for example, what is it actually telling us about reality? Third, and even more importantly, once we revive the specifically ontological perspective, then there is the issue famously made vivid by Quine himself Given that we are to accept a given scientific theory as true, how are we to construe its quantificational structure? Over what entities should our quantifiers range, in the best formulation of the theory in question? So long as any of these three suggestions proves acceptable—provides a nontrivial way to take the question as to how the world has to be, for the scientist's claim that *P* to be true—then the threat of triviality will have been met

Thus we have several strategies for beefing up the excessively modest reading of Quine's program into something more substantial—something with a more active role for philosophy But too much beef would be a bad thing, as we can see by considering a view at the other extreme According to this extreme version of the ambitious reading, the path to metaphysics begins with *Science* and *Nature*—with the

journals, and other data of a similar kind, not with the world For on this view metaphysics relies on a distinctive mode of inference, an inference that takes as input what the scientific experts *say,* and delivers as output conclusions about the nature of reality The behavior of scientists thus becomes *evidence* for a distinctive mode of investigation—the quest for the truth makers, the interpretations, the preferred logical representations, of the claims of science

According to the modest view, then, scientists play all the significant roles in the game of metaphysics Unless philosophers are prepared to become scientists themselves, their role can only be to stand outside—observing the action, perhaps describing it in new and useful terms, but not making the running According to the ambitious view, however, philosophers are investigators in their own right, standing outside the scientific arena, making judgments about reality on the basis of the behavior of the performers within the arena

As I said, this is a caricature What is helpful about it is that it highlights a tension also present in some would-be more accurate readings of Quine's program For imagine the perspective of a scientific expert, seriously concerned to find out how things are According to this caricatured version of the beefy view, her investigation of nature is inevitably a two-stage process First, she needs to develop the best scientific theories about the matter Then she needs to conduct another investigation altogether She needs to step outside the scientific arena—off the scientific stage, onto the philosophical stage—to ask what makes those theories true, how they should be interpreted, what their best logical representation is, or whatever Only from the latter stage can she say, "Ah, so this is how things really are!"

However, this is both absurd, and clearly in conflict with Quine's intentions It is in conflict with Quine because he insists that there is no "outside"—philosophy and science share the same raft, the same stage And it is absurd because if our scientist thinks that she is engaged in the same project within the scientific arena as without, she cannot think *both* that the theories she settles on within the arena are the best that science can do, *and* that there are further questions that can be settled outside (for anything that can be done outside can also be done inside)—while if she thinks she is engaged on different projects, then what reason could she have for thinking that the best theories *inside* are the appropriate input to the investigation *outside*?

To avoid both the absurdity and the conflict with Quine's manifest intentions, then, we need to ensure that in any less caricatured presentation of Quine's program for ontology, ontology is not something that scientists themselves could not regard as continuous with their

own investigations It can be something that, in practice, does not *interest* most working scientists—something that pushes questions they recognize as legitimate further than they themselves feel the need to go But it cannot be a second-order reflection on the results of their investigations

With this in mind, let us look again at Yablo's gloss on Quine's program Under one (perhaps slightly uncharitable) reading, it provides an illustration of the danger that concerns us here—the overly beefy Charybdis, from which any acceptable interpretation of Quine's ontological program needs to distance itself

> Look for the best overall theory—best by ordinary scientific standards *or principled extensions thereof*—and then consider what has to exist for the theory to be true [emphasis added]

We have seen that without absurdity, and without obvious conflict with Quine's insistence that there is no second-order standpoint for ontology, we can indeed make sense of the program thus described, *as it would be without the italicized phrase* The crucial point is that being the best by ordinary scientific standards is compatible with being less than the best by the more refined standards of Quinean ontologists—who, *though scientists at heart themselves,* simply *care more* about truth makers, interpretation, and proper logical form In other words, so long as "best by ordinary scientific standards" is still suboptimal, there is a niche for Quinean ontologists, on the scientific stage But if the investigations conducted by these ontologists are really not different in kind from those of ordinary scientists, they must surely count as principled extensions of ordinary scientific practice To think that there is anything left to do after finding the best overall theory *by those extended standards,* is to lead us back to the trap we have just escaped With the italicized phrase, then, Yablo's gloss on Quine's program suggests the overly ambitious version

Summing up, we have seen that we can make sense of the idea that there is real work for philosophers within the scientific arena, tackling tasks that tax the skills and stamina of scientists themselves In this respect, our caricatured version of the modest reading of Quine's program of ontology is not the only alternative But the new version is nevertheless modest, compared to our caricatured version of the ambitious reading And we need to be on our guard against sliding back in that direction—against glossing the Quinean program in ways which do make ontology second-order, an activity which takes place after and outside science

It seems to me that this overly ambitious reading of Quine's program is actually common in contemporary philosophy As evidence

for this claim, I want to call attention to an ambiguity in presentations of an argument standardly attributed to Quine, the so-called "indispensability argument" for the existence of abstract objects I want to show that this argument has two interpretations, paralleling the modest and ambitious readings of the Quinean program for ontology The stronger interpretation makes the same sorts of mistakes as the version of the ambitious reading described above And yet, as we will see, it is alive and well in the literature

IV IS THERE AN ARGUMENT FROM INDISPENSABILTY?

Realists in the philosophy of mathematics frequently appeal to, and irrealists in the same field frequently seek to evade, what both sides refer to as the Quine-Putnam argument from indispensability What is this argument? Here is a characterization by Field, perhaps the leading contemporary writer on the irrealist side of these debates

> Putnam 1971[12] is the *locus classicus* for the view that we need to regard mathematics as true because only by doing so can we explain the utility of mathematics in other areas for instance, its utility in science and in metalogic The general form of this Putnamian argument is as follows
>
> (i) We need to speak in terms of mathematical entities in doing science, metalogic, etc ,
> (ii) If we need to speak in terms of a kind of entity for such important purposes, we have excellent reason for supposing that that kind of entity exists (or at least, that claims that on their face state the existence of such entities are true) [13]

Here is another recent formulation of the argument, from Mark Colyvan

> For future reference I'll state the Quine-Putnam indispensability argument in the following explicit form
>
> (P1) We ought to have ontological commitment to all and only the entities that are indispensable to our best scientific theories
> (P2) Mathematical entities are indispensable to our best scientific theories
> (C) We ought to have ontological commitment to mathematical entities [14]

[12] Putnam, "Philosophy of Logic," in his *Mathematics, Matter and Method Philosophical Papers, Volume 1* (New York Cambridge, 1979), pp 323–57 (Originally published as *Philosophy of Logic* (New York Harper, 1971))

[13] Field, "Mathematical Objectivity and Mathematical Objects," in his *Truth and the Absence of Fact* (New York Oxford, 2001), pp 315–31, at pp 328–29

[14] Colyvan, "Indispensability Arguments in the Philosophy of Mathematics," in Edward N Zalta, ed , *The Stanford Encyclopedia of Philosophy* (Fall 2003 edition)

In my view, however, the forms of the indispensability argument offered here by Field and Colyvan involve a subtle misinterpretation of Quine and perhaps of Putnam—though admittedly a misinterpretation that neither Quine nor Putnam seems to have done much to discourage Here is Putnam's own version, from the source cited by Field

> So far I have been developing an argument for realism along roughly the following lines quantification over mathematical entities is indispensable for science, both formal and physical, therefore we should accept such quantification, but this commits us to accepting the existence of the mathematical entities in question This type of argument stems, of course, from Quine, who has for years stressed both the indispensability of quantification over mathematical entities and the intellectual dishonesty of denying the existence of what one daily presupposes [15]

I want to call particular attention to Putnam's final remark here—his gloss of Quine if quantification over mathematical entitites is indispensable, it is "intellectually dishonest" to deny the existence of such entities The crucial point—a point missed by Putnam himself here, so far as I can see—is that a principled exclusion of arguments *against* the existence of entities of a certain kind does not in itself comprise an argument *for* the existence of such entities, of the kind supposedly captured by the above formulations

One way to see this is to note that if there were an argument usable by philosophy here, then by Quine's lights it would also be an argument usable by scientists and mathematicians themselves After all, as I have stressed above, Quine insists that philosophy is not separate from science—we are all adrift in the same boat But think about the (supposed) argument as used by scientists themselves To secure premise (P2) (in the notation used by Colyvan above), they must come to accept that quantification over mathematical entities is indispensable—not merely something that they do do, as scientists, but something that survives under reflection—something they think that they do not have a choice about, if they are to continue to do science at all

But for Quine, of course, there is no space between ontological commitment—belief that there are mathematical entities—and acceptance of quantification over mathematical entities So, by Quine's lights, to be in a position to accept (P2) is to accept not only that one believes that there are mathematical entities, but that one is justified in doing so, by the lights of best (philosophically informed) scientific practice It is to believe not only that there are mathematical entities,

[15] Field, "Mathematical Objectivity and Mathematical Objects," p 347

but that one ought to believe that there are (by the standards of scientific practice), having properly considered the alternatives

Imagine our scientists, thus equipped with premise (P2) If they accept premise (P1), they are thus led to the conclusion, (C), that they ought to believe that there are mathematical entities But they believed that already, by assumption, if 'ought' means something like 'by the internal standards of science' So the argument could only take them somewhere new if there were some other standards—some other standpoint, from which to assess the question as to whether there are mathematical entities

As before, there are two problems with this last idea (that is, that there is some other standpoint from which to assess the question) One is that it flatly contradicts Quine, who insists that there is no separate standpoint for ontology, outside that of science The other is that by introducing two standards for ontological commitment—the second-rate "as-if" kind of commitment at the first stage, as compared to the first-rate, meaty kind of commitment at the second—it pulls the rug from beneath the entire argument If there is a second-rate kind of ontological commitment, why should *that* kind of commitment be a guide to what there is? On the contrary, presumably, what makes it second-rate is that it is not a (first-rate) guide to what there is

Thus the argument from indispensability seems to embody the beefy vices of our excessively ambitious interpretation of Quinean ontology If so, then our Charybdis was not a caricature, after all, except in flaunting what proponents of the argument from indispensability have disguised, the fact that there can only be such an argument if the standpoint of ontology is not merely that of science

In defense of the argument from indispensability, it might be said that Quine insists that if science reaches that stage of accepting (P2), then there is no philosophical standpoint from which it makes sense to doubt that there are mathematical entities—to ask "But are there REALLY mathematical entities?" Doesn't this imply that if science reaches the stage of accepting (P2), then we are justified in affirming that there are mathematical entities—after all, aren't we justified in affirming what it makes no sense to doubt?

Well, it depends Perhaps we are justified in repeating what science says (though the issue of the source of the normative force of 'justified' here is a nice one—more on this below) But even if so, this involves no inference from the fact that science says it no argument, simply concurrence The Quinean doctrine that if science reaches that stage of accepting (P2), then there is no philosophical standpoint from which it makes sense to doubt that there are mathematical entities—to ask "But are there REALLY mathematical entities?"—does

put paid to a certain sort of ontological skepticism, or anti-realism But it does not imply that there is an argument *from* the needs of science *to* ontological conclusions—*for* realism On the contrary, it deflates or disallows a certain sort of ontological debate a debate taking place outside science, about whether there are things of the kind science quantifies over After all, think of 'REALLY' as a metaphysician's term of art The argument that it makes no sense to ask "But are there REALLY mathematical entities?" does not imply that we should say "There REALLY are mathematical entities " Perhaps we should simply forget about 'REALLY'

The difficulty with the argument just given is that our realist opponents will deny that they ever meant anything special (namely, 'REALLY') by 'really' A familiar dispute then ensues about whose position is the more modest—about who holds the metaphysical low ground, so to speak From the deflationist's point of view, the right strategy is to present one's opponent with issues on which she must take a stand, one way or the other The aim is to show that if she agrees, she is being more deflationist than she wants to be, while if she disagrees, she holds commitments sufficiently inflated to be targets The claimed argument from indispensability provides one such choice point, in my view However, if someone insists that she meant the argument only in the modest, anti-realist dismissing sense, then we deflationists have no reason to argue with her any further On the contrary, we should welcome her to the anti-metaphysical club—to that enlightened circle who agree with Carnap, in rejecting "both the thesis of the reality of the external world and the thesis of its irreality "

Summing up, we have seen that the only acceptable Quinean ontology is a very modest variety It need not be as modest as a mere chorus line to science There seems to be more dignified employment for philosophy, dealing with issues that, although continuous with science, suit the skills and temperament of philosophers better than those of working scientists themselves But it is not an activity "after" science, in a logical sense In particular, it does not rely on inferences from what best current science says about reality On the contrary, it is itself a kind of formal finishing school for scientific theory Its own products *are* best current science

This conception of the role of philosophical ontology is certainly modest, compared to some alternatives However, I think it is worth noting in passing that it may still seem somewhat self-important, compared to views genuinely deferential to science After all (it might be objected), how could there really be space for such a pursuit? If scientists themselves get by without doing it, it cannot really be indispensable to science—in which case, what rationale can it have, for

theorists whose role is as aides to science, if anything? Quinean ontology walks a rather fine line here, I think

However, my present interest is not how much role there really is for ontology *within* science It is the question as to what basis, if any, there can be for a distinction relevant to ontology *between* science and other areas of discourse Let us return to that issue, in the light of the conclusion that Quinean ontology is necessarily a very modest, deflationary kind of ontology

V NATURALISM THICK AND THIN

In its ambitious form, the argument from indispensability exemplifies a proposal for doing metaphysics from outside science—a proposal that metaphysics should appeal to the practices of science as *evidence*[16] for its own, separate, investigation Note the special role accorded to *science,* of all possible intellectual pursuits The strong form of the indispensability argument does not accord any weight to the ontological commitments of moralists or musicians, say, or marketeers, or even mathematicians themselves, except in so far as those commitments meet the needs of scientists

If this preference for science could be justified—without begging the question, of course—then it would give us a basis for scientific naturalism, within the confines of this ambitious, non-Quinean metaphysical program Ambitious ontology would have turned out to be a discipline which, in virtue of special place of the needs of science in its evidential base, does yield the conclusion that what exists is what (mature) science reveals But the prospects for such a justification are surely slim The evidential principle simply assumes naturalism So long as naturalism is in doubt, so too will be the proper evidential base for this kind of ontology by inference from best first-order practice

However, even this slim prospect disappears when we move to the modest program In this context, the relevance of argument that you cannot do science without committing yourself to numbers is much like that of the argument that you cannot make an omelette without breaking eggs It leaves you with a choice commit to numbers or do not do science, break eggs or do not make omelettes

So far as I can see, it does not supply any normative pressure to choose the former option Someone already committed to science has no choice, of course—it is from that internal, already-committed perspective that the argument has some normative force But modest ontology can exert no force on mere browsers—intellectual window-

[16] Where 'evidence' means more than testimony, of course—that way modesty lies

shoppers, simply interested in what the package entails, what the recipe requires Even more importantly, for present purposes, there is no sense in which choice *privileges* science On the contrary, it is clearly possible that there will turn out to be other activities such that you cannot do X without postulating Y What is not admissible, it seems, is that there could be a perspective from which this possibility could be ruled out If so, then there can be no a priori argument for naturalism We are naturalists, de facto, if it turns out that all we are committed to is what is required by science But even if that is so, it is not a position we reach by argument

The crucial point is that the restriction of ontology to first-order ontology—the rejection, with Quine, of any higher standpoint for philosophy—does not necessarily amount to a restriction of ontology to first-order *scientific* ontology It all depends on whether science is the only existentially quantifying game in town, at first-order level As I have already noted, Carnap takes the negative side on this issue in "Empiricism, Semantics, and Ontology," defending a pluralist conception of the role of existential quantification in language

However, one component of Quine's famous criticism of Carnap focuses precisely on this issue, and Quine is also known as a critic of other manifestations of pluralism about existence and existential quantification, notably that of Ryle I want to show that these Quinean arguments contain little to trouble Carnap's combination of deflationism about metaphysics and pluralism about the functions of linguistic categories As a result, they provide no serious obstacle to the suggestion that in virtue of such pluralism, not all first-order ontological commitment need be scientific ontological commitment

Quine's objections to Carnap on this matter also offer an apparent defense of metaphysics against Carnap's criticisms—a defense in tension, it may seem, with my suggestion that Quine, too, is a deflationist about ontological issues Before turning to the issue of pluralism, I want to show that in fact there is no tension here For all practical purposes, Quine does agree with Carnap about the status of metaphysical issues If anything, he is more of a pragmatist than Carnap, arguing that Carnap is mistaken in assigning a more robust status to scientific matters [17]

VI QUINE'S DEFENSE OF METAPHYSICS—THE BAD NEWS

Much of Quine's attack on Carnap turns on the objection that Carnap's notion of a linguistic framework presupposes the analytic–synthetic

[17] For a more detailed treatment of these matters, see my "Carnap, Quine, and the Fate of Metaphysics," *Electronic Journal of Analytic Philosophy*, Issue 5 (Spring 1997) The next two sections draw extensively on the discussion in that paper

distinction Quine argues that in virtue of the failure of the analytic–synthetic distinction, even internal question are ultimately pragmatic Referring to Carnap's view that, as Quine puts it, "philosophical questions are only apparently about sorts of objects, and are really pragmatic questions of language policy," Quine asks "But why should this be true of the philosophical questions and not of theoretical questions generally? Such a distinction of status is of a piece with the notion of analyticity, and as little to be trusted "[18] In other words, Quine's claim is that there are no purely internal issues, in Carnap's sense No issue is ever entirely insulated from pragmatic concerns about the possible effects of revisions of the framework itself Pragmatic issues of this kind are always on the agenda, at least implicitly In the last analysis, all judgments are pragmatic in nature

Grant that this is true What effect does it have on Carnap's antimetaphysical conclusions? Carnap's internal issues were of no use to traditional metaphysics, and metaphysics does not lose if they are disallowed But does it gain? Science and mathematics certainly lose, in the sense that they become less pure, more pragmatic, but this is not a gain for metaphysics And Quine's move certainly does not restore the nonpragmatic external perspective required by metaphysics In effect, the traditional metaphysician wants to be able to say, "I agree it is useful to say this, but is it true?" Carnap rules out this question, and Quine does not rule it back in [19]

Quine sometimes invites confusion on this point He says that

> if there is no proper distinction between analytic and synthetic, then no basis at all remains for the contrast which Carnap urges between ontological statements [that is, the metaphysical statements that Carnap wants to disallow] and empirical statements of existence Ontological questions then end up on a par with the questions of natural science [20]

This sounds like good news for ontology, but actually it is not Quine's criticism of Carnap cannot provide vindication of traditional metaphysics, for if all issues are ultimately pragmatic, there cannot be the more-than-pragmatic issue of the kind the metaphysician requires The main effect of abandoning the analytic–synthetic distinction is that Carnap's distinctions are no longer sharp—there are no purely

[18] Quine, *Word and Object* (Cambridge MIT, 1960), p 271

[19] Roughly, Carnap allows us to ask about truth only for internal questions Quine agrees, but says that there are no such questions, in the last analysis, because there are no firm linguistic rules

[20] Quine, "On Carnap's Views on Ontology," in *The Ways of Paradox and Other Essays* (New York Random House, 1966), pp 126–34, at p 134

internal (nonpragmatic) issues, because linguistic rules are never absolute, and pragmatic restructuring is never entirely off the agenda But a metaphysician who takes this as a vindication of his position—who announces triumphantly that Quine has shown us that metaphysics is in the same boat as natural science, that "ontological questions [are] on a par with the questions of natural science"—is someone who has not been told the terrible news Quine himself has sunk the metaphysicians' traditional boat, and left all of us, scientists and ontologists, clinging to Neurath's Raft [21]

As Quine himself puts it in the same piece

> Carnap maintains that ontological questions are questions not of fact but of choosing a convenient scheme or framework for science, and with this I agree only if the same be conceded for every scientific hypothesis [22]

Thus Quine is not returning to the kind of metaphysics rejected by the logical empiricists On the contrary, he is moving forwards, embracing a more thoroughgoing post-positivist pragmatism In this respect, far from blocking Carnap's drive towards a more pragmatic, less metaphysical destination, Quine simply overtakes him, and pushes further in the same direction

It might be objected that news still looks much better for metaphysics than Carnap would have had us believe Granted, there is no longer any pure, nonpragmatic science to be had, and no nonpragmatic metaphysics, either But if metaphysics nevertheless ends up "on a par" with the kinds of questions investigated at CERN and Bell Labs, is not that a kind of respectability worth having?

However, this objection trades on an excessively optimistic reading of the phrase 'on a par' Like the rejection of the analytic–synthetic distinction on which it depends, the effect of Quine's challenge is simply to soften the distinction that Carnap wants to draw between ontological issues that are settled "by convention," when we adopt a framework, and ontological issues that are serious matters for investigation, within a framework In effect, Quine turns a dichotomy into a gradation, and lops off the pure cases, at either end At one end, as I have noted, this makes science a little more pragmatic At the other end, it makes the issue as to whether there are numbers, or

[21] The immodest interpretation noted above would have us do ontology by studying the behavior of the creatures on the raft from the outside, as it were, but that interpretation is absurd

[22] Quine, "On Carnap's View on Ontology," p 134 Note Quine's revealing use of the phrase 'for science' It is far from clear that for Carnap, the convenience of adopting a linguistic framework is always convenience *for science*

physical objects, just a tiny bit less trivial—remember, no "quite black threads [nor] any white ones." Patently, however, this change in logical status is not enough to support a serious metaphysical investigation into the status of numbers, or physical objects—no more than the same news supports a serious sociological investigation into the gender and marital status of bachelors.

Similarly, the news that science is ultimately pragmatic does not mean that CERN and Bell Labs should be hiring pragmatists. There is still a big difference, in practice, between the day-to-day business of empirical science and the sort of rare occasions on which Quinean science has to confront its pragmatic foundations. At best, it is with these rare situations that Quine's response to Carnap can compare metaphysics—and patently, they are no serious challenge to Carnap's objections to traditional metaphysics. Once again, the force of Quine's remarks is not that metaphysics is like science as traditionally (that is, nonpragmatically) conceived, but that science (at least potentially, and at least in extremis) is like metaphysics as pragmatically conceived.

VII. QUINE'S OBJECTIONS TO PLURALISM

As I have stressed, however, there is also a respect in which Quine does not move in the same direction as Carnap. Think of Carnap's move as a vector sum of two components, the first a push towards a deflationary and pragmatic conception of ontological issues, the second a push towards pluralism. Quine goes further than Carnap on the first axis, but resists all movement on the second. What, then, are Quine's objections to Carnap's pluralism?

I have already noted that Carnap's pluralism operates at two levels. On the surface, most explicitly, it is a doctrine expressed in terms of the logical syntax of language—the view that language may be significantly factored into distinct linguistic frameworks, each associated with "a particular style of bound variables," as Quine puts it.[23] Underlying this logico-syntactical pluralism, however, is the pragmatic or functional pluralism which provides its motivation. Carnap holds that there is some sort of category mistake involved in assimilating issues of the existence of classes, say, and the existence of physical objects. His model for the construction of linguistic frameworks reflects this assumption, requiring that we mark the category boundaries in our choice of syntax—a different quantifier for each category, for example. But the distinctions in question are not grounded at the syntactical level.

This is important, because Quine's challenge to Carnap's pluralism rests on a challenge to its logico-syntactical manifestation. Quine

[23] Quine, "On Carnap's View on Ontology," p. 130.

argues that it cannot be more than "a rather trivial consideration" whether we use different quantifiers for numbers, classes, and physical objects, for example, or use a single existential quantifier ranging over entities of any of these kinds But it seems to me that we can allow that Quine is right about this, while insisting that it makes no difference at all to the issue that really matters namely, whether Carnap is right about the underlying functional distinctions, and right about category mistakes

The notion of a category mistake was familiar to the logical positivists of the 1920s and 1930s In the *Aufbau* of 1928, Carnap himself uses the term 'mixing of spheres' *(Spharenvermengung)* for, as he puts it later, "the neglect of distinctions in the logical types of various kinds of concepts "[24] But for contemporary audiences the notion is particularly associated with Ryle Ryle is quite clear that it has implications for ontological issues, and in a famous passage in *The Concept of Mind,* touches on the question as to whether existence is a univocal notion

> It is perfectly proper to say, in one logical tone of voice, that there exist minds, and to say, in another logical tone of voice, that there exist bodies But these expressions do not indicate two different species of existence, for "existence" is not a generic word like "colored" or "sexed " They indicate two different senses of "exist," somewhat as "rising" has different senses in "the tide is rising," "hopes are rising" and "the average age of death is rising " A man would be thought to be making a poor joke who said that three things are now rising, namely the tide, hopes and the average age of death It would be just as good or bad a joke to say that there exist prime numbers and Wednesdays and public opinions and navies, or that there exist both minds and bodies [25]

Given Quine's response to Carnap, it is not surprising that he has little sympathy for Ryle's apparent ontological pluralism In a section of *Word and Object* devoted to ambiguity, Quine takes the opportunity to put on record his objection to Ryle's view

> There are philosophers who stoutly maintain that "true" said of logical or mathematical laws and "true" said of weather predictions or suspects' confessions are two uses of an ambiguous term "true " There are philosophers who stoutly maintain that "exists" said of numbers, classes and the like and "exists" said of material objects are two uses of an ambiguous term "exists " What mainly baffles me is the stoutness of their maintenance What can they possibly count as evidence? Why not view "true" as

[24] P Schilpp, ed , *The Philosophy of Rudolf Carnap* (La Salle, IL Open Court, 1963), p 45

[25] Ryle, *The Concept of Mind* (London Hutchinson, 1949), at p 23

> unambiguous but very general, and recognize the difference between true logical laws and true confessions as a difference merely between logical laws and confessions? And correspondingly for existence?[26]

But what is the disagreement between Quine and Ryle? For Quine, matters of ontology reduce to matters of quantification, and presumably Ryle would not deny that we should quantify over prime numbers, days of the week, and dispositions Indeed, Ryle might reinforce his own denial that there are "two species of existence" by agreeing with Quine that what is essential to the single species of existence is its link with quantification Ryle simply needs to say that what we are doing in saying that beliefs exist is not what we are doing in saying that tables exist—but that this difference rests on a difference in talk about tables and talk about beliefs, rather than on any difference in the notions of existence involved So far this is exactly what Quine would have us say The difference is that whereas Quine's formulation might lead us to focus on the issue of the difference between tables and beliefs per se, Ryle's functional orientation—his attention to the question as to what a linguistic category *does*—will instead lead us to focus on the difference between the *functions* of talk of beliefs and talk of tables

Moreover, it is open to Ryle (and again, entirely in keeping with his use of the analogy with 'rising') to say that in one important sense, it is *exactly the same* existential quantifier we use in these different cases It is the same logical device, but employed in the service of different functional, pragmatic, or linguistic ends This move is important because it goes a long way to defusing Quine's objection to Carnap

By way of comparison, consider the familiar view that the truth predicate is a grammatical device to meet certain logical and pragmatic needs a device for disquotational or prosentential purposes, say As a number of writers have noted,[27] this account is compatible with the view that declarative sentences can perform radically different functions, in a way which is not captured merely by noting differences in content Consider projectivism about moral or causal claims, for example A deflationist may say that although it is the same deflated notion of truth we use when we say there are moral truths, or that there are causal truths, moral and causal claims have quite different functions (both with respect to each other, and with respect to other kinds of declarative claims)

[26] Quine, *Word and Object*, p 131 The above passage from *The Concept of Mind* is one of two places to which Quine refers readers for "examples of what I am protesting"

[27] See, for example, Horwich, *Truth* (New York Basil Blackwell, 1990), at pp 87–88, Blackburn, *Spreading the Word* (New York Oxford, 1984)

An analogous move seems to provide the best way to preserve the pluralist insights of Carnap and Ryle in the face of Quine's objections We should concede to Quine that there is a single device of existential quantification, just as there is a single device of disquotational truth—if Carnap was really committed to the view that there are different existential quantifier, one for each framework, then he was wrong about that [28] But we should insist that this device has application in a range of cases, whose functional origins are sufficiently distinct that naturalism is guilty of a serious error, in attempting to treat them as all on a par (Compare There is just one device of enumeration by units, but it gives rise to incommensurable notions of 'rising', when the units in question belong to different scales, serving different purposes)

A good way to put this might be to say that the subject-predicate form, and indeed the notion of an object itself, have this one-many functional character In one sense, it is the same tool or set of tools we employ wherever we speak of objects, or whenever we use the subject-predicate form, or—what seems part of the same package—whenever we use the existential quantifier However, there is no further unitary notion of *object,* or *substance,* or metaphysical bearer of properties, but "only a subject position in an infinite web of discourses "[29] Similarly, it is the same tool or set of tools we use whenever we speak of truth, whenever we make a judgment or an assertion But in each case, the relevant tool or set of tools may have incommensurable uses, if there are important senses in which the bits of language they facilitate have different functions (in a way which does not simply collapse into differences in the objects *talked about*)

[28] Though it is hard to see that there could really be a substantial difference of opinion here We could index our disquotational truth predicates in a way which distinguished the predicate we apply to moral claims from the predicate we apply to causal claims, but this trivial syntactical exercise would not prevent it from being the case that the resulting predicates both serve the same disquotational function It is surely uncharitable to Carnap to suggest that he was confused about the analogous point, in the case of the existential quantifier A champion of less deflationary metaphysics might think that there were significant distinctions for such syntactical conventions to mark, but why should Carnap think so?

[29] To reverse the sense of a remark by one of David Lodge's characters, who is characterizing the view that there is no such thing as the Self In this context, I note that Putnam does want to distinguish between "speaking of objects" and "using the existential quantifier," and wants to use the term object in a more restricted sense (See his *Ethics without Ontology,* pp 52ff , and "Was Wittgenstein *Really* an Anti-Realist about Mathematics?" in Timothy McCarthy and Sean C Stidd, eds *Wittgenstein in America* (New York Oxford, 2001), pp 140–94) However, there does not seem to be much at issue here Certainly, the Carnapian view I am recommending seems close to Putnam's "pragmatic pluralism" (see *Ethics without Ontology,* pp 21–22)

This line of thought leads to an explanatory issue To what do we need to appeal, in explaining the difference between talk of chairs and talk of beliefs? If it is simply to difference between chairs and beliefs, pluralism has made no headway An interesting pluralism stems from reversing the order of explanation explaining differences between our talk of different kinds of objects in terms of differences in the function of the talk, not differences in the objects Thus a projectivist about values and causes explains the differences between evaluative and causal judgments in terms, ultimately, of functional differences between the kinds of psychological states projected in the two cases desires, perhaps, in the case of evaluative judgments, and Humean habits of expectation, in the case of causal judgments

This explanatory program leads us in turn in the direction of Eleatic naturalism In one common version, the Eleatic criterion for realism is that the entities in question figure in causal explanations of our beliefs and talk (apparently) about those entities To finish, I want to turn briefly to the question whether the Eleatic form of naturalism can hold the line, where Quinean naturalism could not, against Carnap's pale but permissive pluralism First, however, let me summarize the case against Quinean naturalism

I have argued that Quine's appeal to the virtues of a single existential quantifier does not undermine Carnap's functional pluralism Given a sufficiently deflated view of ontology, functional pluralism does not depend on *syntactical* pluralism about existential quantification Hence it remains open to a Carnapian to dispute the claim that all ontological commitment is scientific in nature—to defend the view that ontologically committed language may do other jobs Naturalism thus remains vulnerable to the Carnapian challenge

But suppose Quine had won this dispute Suppose that he had established that in some substantial sense, existential quantification is a more monochrome matter than Carnap takes it to be Would this have been a victory for naturalism? Not at all, it seems to me It is one thing to establish that all ontological issues are the same kind of issue, quite another to establish that they are all *scientific* issues So victory on this point would not have vindicated Quinean naturalism It would simply have blocked what is otherwise a powerful objection to Quinean naturalism, namely, the thesis that the relevant parts of language are functionally pluralistic, so that naturalism is guilty of a category mistake [30] But although functional pluralism defeats naturalism, functional monism does not imply naturalism

[30] Note that this objection derives its power, in large measure, from the fact that it comes from within science—from a scientific reflection on what we humans do with our

VIII ELEATIC NATURALISM TO THE RESCUE?

Now to Eleatic naturalism As we noted earlier, one appealing version of the Eleatic criterion holds that we should be realists about a class of entities—about *X*s, let us say—when *X*s figure in causal explanations of our talk and beliefs "about" *X*s (that is, of talk which has the superficial form of talk about *X*s) Simon Blackburn canvasses such a criterion as the basis of a distinction between quasi-realism and genuine realism, for example

> A quasi-realist [for example, about value] can mimic our formal practice with the concept of truth or fact But surely he cannot give the facts any role in explaining our practice To do so is to embrace their real distinct existence, or so it might seem [31]

Can such a principle hold the line, where Quinean naturalism could not, against Carnap's pale but permissive pluralism (or indeed against Quine's own metaphysical deflationism)?

The basic difficulty with this suggestion is that by the deflationist's lights, the Eleatic view simply has an overly inflated conception of what is at issue After all, suppose it turns out that we find a pragmatic need for existential quantification in cases which do not pass the Eleatic test To regard the Eleatic principle as an *ontological* criterion is to introduce a distinction between "weak" ontological commitment, for which existential quantification alone is sufficient, and "strong" or "genuine" commitment, for which the Eleatic criterion is the benchmark But such a distinction simply flies in the face of deflationism Whence, and what, this stronger notion of ontological commitment, for which causation is supposed to be the *sine qua non*?

This point may have been been obscured by the fact that in one sense, Quine himself attempts to draw such a distinction, favoring the ontological commitments required by science over what he treats as second-rate commitments Quine's proposal challenges the Eleatic criterion, of course, as we noted at the beginning, and this challenge

---

language Hence it points to a tension between two strands in Quine's naturalism His own insistence that human thought and talk be addressed from a scientific perspective threatens to undermine his assumption that all serious ontological commitment is in the service of science itself All it takes, in effect, is for science of linguistic behavior to reveal that we humans do other things with existential quantifiers, besides scientific theorizing

[31] Blackburn,"Truth, Realism, and the Regulation of Theory," in *Essays in Quasi-Realism* (New York Oxford, 1993), pp 15–34, at p 31 (Cf *Spreading the Word,* p 257) Blackburn goes on to call attention to some difficulties for this suggestion, however—here, as elsewhere, he is by no means an unqualified Eleatic naturalist

may have obscured the more basic point if we take Quine's deflationism seriously, the serious challenger to Eleatic naturalism is not Quinean naturalism—another high-contrast view, but adjusted to accommodate numbers, abstract objects, and the like—but a paler and much more inclusive deflationary metaphysics In other words, the more serious challenge stems from the fact that Quine's own account of what is at stake in ontological matters leaves no apparent space for a distinction between first-rate, "genuine," ontological commitment, and lesser varieties of existential quantification [32]

Thus there seems to be a fundamental tension between Eleatic naturalism and the Quinean message for metaphysics, properly (that is, "thinly") construed Quinean deflationism counts against Eleatic naturalism, as much as its "Quinean" cousin

So much the worse for Quinean deflationism, perhaps—one can certainly imagine an Eleatic challenge to the Quinean message Such a challenge would not conflict with the main claim of this paper, namely, that accepting the Quinean view means rejecting naturalism, in either form But I want to finish by noting something that a Carnapian pluralist can do to defuse the kind of intuition that seems to support the Eleatic proposal

IX THE PERSPECTIVAL FALLACY

Recall Blackburn's gloss on the Eleatic intuition to "give the facts [a] role in explaining our practice is to embrace their real distinct existence, or so it might seem " When we seek the causes of our beliefs and practices, we are engaged in forensic science The task of explaining our beliefs and utterances is a small but significant part of the scientific enterprise as a whole

As we engage in this project, there is an inevitable but potentially misleading difference between the scientific framework and others We are engaged in a scientific practice, in seeking the explanations for various other practices (or, indeed, for scientific practice itself) In other words, we are *employing* scientific vocabulary, to *talk about* the genealogy of vocabularies in general We *use* scientific vocabulary, but *mention* the various object vocabularies with which we are concerned

In general, presumably, the explanations we offer from this perspective will appeal to extra-linguistic states of affairs of various kinds—to the various features of ourselves and our environments that explain our linguistic practices Thus, roughly, it is characteristic of the project that it appeals to nonlinguistic ontology, in the service

[32] Yablo argues convincingly for this conclusion, in "Does Ontology Rest on a Mistake?"

of explanations of various kinds of linguistic behavior In the case in which the latter behavior "involves" a distinct ontology of its own—an ontology of moral values, in the moral case, for example—the project embodies the starkest possible asymmetry between this ontology and that of science It invokes scientific ontology, while ignoring moral ontology (Our *explanandum* is the use of moral language, not moral states of affairs)

No wonder, then, that the natural facts that play a role in explaining our practices *look* privileged from this perspective They are privileged, *from this perspective,* for it simply is the scientific explanatory perspective However, if we take this perspectival privilege for an absolute ontological criterion, we must be presupposing that science is (necessarily) the only ontologically committing game in town I have suggested that science itself ought to challenge that presupposition, regarding the range, functions, and possible plurality of our ontological commitments to be a matter to be addressed a posteriori, from within a naturalistic reflection on our own linguistic behavior

If this seems doubtful, note that we can consider our linguistic practices from other perspectives We can evaluate them, in various senses, for example [33] If we invoke evaluative or normative properties in this context, the resulting ontological commitment is once again a product of the perspective—a product of the framework in play, in Carnapian terms (Again, none of this depends on a rigid syntactical partition of frameworks, of the kind opposed by Quine)

Thus Carnapian pluralism offers a natural (indeed, naturalistically respectable) way to deflate the Eleatic intuition—to explain it as a product of a kind of perspectival fallacy We Carnapians should certainly embrace the project of explaining our linguistic practices—for that way, if all goes well, lies a scientific foundation for the suggestion that different parts of language serve different functional ends, in some sense overlooked by Quinean and Eleatic naturalists As I have noted elsewhere,[34] the upshot would be that science might properly take a more modest view of its own importance Naturalism of both varieties would be defeated from within, as it were, by a scientific discovery that science is just one thing among many that we do with the linguistic tools of ontological commitment

[33] Arguably, in fact, they cannot count as full-blown linguistic practices—as "sayings," or "assertings," say—unless they are taken to be subject to normative assessments of various kinds But the present point does not depend on this claim

[34] See my "Naturalism without Representationalism," in David Macarthur and Mario de Caro, eds, *Naturalism in Question* (Cambridge Harvard, 2004), pp 71–88

For the moment, pending such developments, Carnapian pluralism remains somewhat promissory I have argued, however, that it has nothing to fear from the Eleatic challenge, or from Quine's objections to Carnap's own formulation of the pluralist view In all respects bar one, in fact, Quine is an ally Most importantly, he, too, is a metaphysical deflationist—thereby endorsing a view that already drains most of the blood from the issue between naturalists and their opponents The one significant disagreement between Carnap and Quine turns on Quine's view that science is the only game in town—the only serious activity to make use of existential quantification And on this matter, as I have said, we Carnapians should put our faith in science itself

HUW PRICE

University of Sydney

SAFETY, CONTENT, APRIORITY, SELF-KNOWLEDGE*

This essay motivates a revised version of the epistemic condition of *safety* and then employs the revision to (i) challenge traditional conceptions of apriority, (ii) refute "strong privileged access," and (iii) resolve a well-known puzzle about externalism and self-knowledge

I VARIETIES OF EPISTEMIC RISK

I think I hear a lark outside Suppose I am right, but my belief could easily have gone amiss In that case I do not know that I hear a lark For short if I know that *p*, I must "safely" believe that *p* But how might a belief go amiss? Let me count the ways

First, I might have falsely believed that I hear a lark If this possibility threatens—for example, there are many lark-imitating imposters nearby—then I do not know that I hear a lark As safety is usually formulated, it only rules out this type of threat

STANDARD SAFETY *S* could not easily have falsely believed that *p* [1]

But safety should do more A second way I could have messed up is by falsely believing some other, closely related proposition [2] Suppose I have a true demonstrative thought, one that I might express by saying, 'That is a lark' If the imposters are nearby, I am still in danger

* Many of the ideas in this paper, especially in section III, grew out of conversations with John Hawthorne I owe a great deal to him Warm thanks also to Ryan Wasserman, Brian Weatherson, and Takashi Yagisawa for discussion and comments on earlier versions of this paper

[1] For a sampling of those who endorse some version of safety as a necessary condition on knowledge, see Stewart Cohen, "Contextualism and Skepticism," *Philosophical Issues*, X (2000) 94–107, Keith DeRose, "Solving the Skeptical Problem," *Philosophical Review*, CIV (1995) 1–52, R.M Sainsbury, "Easy Possibilities," *Philosophy and Phenomenological Research*, LVII (1997) 907–19, Ernest Sosa, "Relevant Alternatives, Contextualism Included," *Philosophical Studies*, CXIX (2004) 35–65, and Timothy Williamson, *Knowledge and Its Limits* (New York Oxford, 2000)

[2] Sainsbury, *op cit*, and Brian Weatherson, "Luminous Margins," *The Australasian Journal of Philosophy*, LXXXII (2004) 373–83, make a persuasive case that safety should not merely rule out close worlds where the very same proposition is believed falsely And such a version of safety is required by the epistemic theory of vagueness in Williamson, *Vagueness* (New York Routledge, 1994) For other excellent discussions of safety and content-switching involving demonstratives, see also Jessica Brown, "Reliabilism, Knowledge, and Mental Content," *Proceedings of the Aristotelian Society*, C (2000) 115–35, and James Pryor, "Comments on Sosa's 'Relevant Alternatives, Contextualism Included'," *Philosophical Studies*, CXIX (2004) 67–72

0022-362X/07/0408/403–23 © 2007 The Journal of Philosophy, Inc.

UNIVERSITY OF CALCUTTA ORE CAMPUS

of messing up But the problem is not that I could easily have falsely believed that very proposition (Arguably it is a necessary truth [3]) Had an imposter been singing, I would have believed something else—a proposition with different truth conditions—though I would have expressed my belief the same way Since standard safety only rules out close possibilities in which I falsely believe that *p*, the demonstrative belief counts as safe even though the general belief that I hear some lark or other does not But both beliefs intuitively fail to count as knowledge After all, it is not an epistemically risky undertaking to infer the general truth from the demonstrative one

Yet a third way I might have messed up is by having a thought with gappy content, paradoxical content, or no content at all [4] Suppose that (unwittingly) I often hallucinate lark calls In between some hallucinations, I hear a real lark and form the demonstrative thought I express by 'That bird is a lark' According to a leading semantic picture, if I had been hallucinating instead, my thought episode would have had no content, because only a singular proposition would do In that case, I would have been messing up, even though I would not have believed the proposition I actually do, and the thought I give voice to by 'That bird is a lark' would not have been false (The same point can be illustrated by assuming semantic externalism for predicates [5])

An alternative view has it that, had I been hallucinating, my thought would have had *gappy* content If gappy propositions have no truth value,[6] then intuitively one messes up by believing them we

[3] I assume that the content of a demonstrative thought is a proposition involving some object *a* whose truth value with respect to any possible world depends solely on how things stand with *a* at that world This leaves it open whether, in a successful case, the semantic value of a demonstrative is simply its referent One can achieve the relevant kind of object dependence with a view along the lines of John McDowell, "De Re Senses," *Philosophical Quarterly*, XXXVI (1984) 281–94, or even with a quantificational approach to demonstratives see Jeffrey C King, *Complex Demonstratives* (Cambridge MIT, 2001), and Hawthorne and Manley, *Something in Mind* (New York Oxford, forthcoming), chapter 6

[4] Brown, *op cit*, notes that "local reliability" (in no nearby situation is *p* false and *S*'s method produces the belief that *p*) allows demonstrative thoughts to be known even in cases where the subject is prone to empty thoughts

[5] Suppose that experts settle which precise property we express by 'lark' And suppose that they nearly adopted a more exclusive membership condition, one according to which the bird I hear would not have fallen under the extension of 'lark' Then the belief I would have expressed with 'That is a lark' would have been a false belief involving a closely related property that the demonstrated bird does not have (Here nothing special hinges on the fact that 'lark' is a natural kind term semantic deference is a source of externalism even for functional kind terms for example, 'diode' and 'Wankel rotary engine')

[6] See Nathan Salmon, "Nonexistence," *Noûs*, XXXII (1998) 277–319

aim at truth with our beliefs If instead we consider all atomic gappy propositions to be false and their negations true,[7] we should hold that a subject messes up by believing even *true* gappy propositions—and safety will need revision to account for that fact For I am failing equally, from an epistemic point of view, whether I point at a dark empty corner and say 'He is fat', or I point at a dark empty corner and say 'He is not fat' Supposing (contrary to intuition) that the second sentence counts as true, we should still consider the associated demonstrative thought to be an epistemic failure of the sort whose nearness and similarity undermines knowledge In short, the danger of emptiness—no less than the danger of falsehood—dispels knowledge

This can be illustrated without appealing to object-dependence or social externalism, consider instead the phenomenon of paradox Suppose Mary believes that what Susan just wrote in her diary is false Aside from this belief, all of Mary's beliefs about Susan are true It turns out that Susan wrote 'Everything Mary believes about me is true' As things stand, Mary's belief is paradoxical we cannot evaluate it for truth or falsehood without contradiction But had Susan written 'I had eggs for breakfast' instead, Mary's belief would have been truth evaluable This is the sort of case about which Saul Kripke wrote "Our statements involving the notion of truth [are] risky they risk being paradoxical if the empirical facts are extremely (and unexpectedly) unfavorable There can be no syntactic or semantic 'sieve' that will winnow out the 'bad' cases while preserving the 'good' ones "[8] Whether we say that Mary's thought in the bad case fails to have propositional content, or that its content simply cannot be evaluated for truth,[9] having a paradoxical thought is a way of messing up

Clearly, thoughts can be in imminent danger of paradox while not being in any danger of falsehood suppose Susan actually wrote something false, had no inclination to write anything true, but nearly wrote something paradoxical Then Mary's belief is true, and she is not in danger of having a false belief, but she is in danger of epistemic failure There are even beliefs that can be paradoxical but cannot be untrue for example, the belief that someone is thinking something

[7] This is the view of David Braun, "Empty Names," *Noûs*, XXVII (1993) 449–69 Note that on this approach the gappy proposition expressed by 'Vulcan is not Vulcan' is true, moreover it is the same proposition as that expressed by 'Vulcan is not Santa Claus' See Ben Caplan, "Empty Names," in *The Encyclopedia of Language and Linguistics* (New York Elsevier, 2006, 2nd ed), pp 132–36, and also Anthony Everett, "Empty Names and 'Gappy' Propositions," *Philosophical Studies*, CXVI (2003) 1–36

[8] Kripke, "Outline of a Theory of Truth," this JOURNAL, LXXII, 19 (November 6, 1975) 690–716

[9] As recommended by, for example, Scott Soames, *Understanding Truth* (New York Oxford, 1999)

untrue right now [10] (If someone else is thinking something untrue, then it is true If not, it is paradoxical) It should not follow that this belief is always safe Suppose someone tokens it in a very sparsely populated world in which people spend much of their time in deep sleep If a belief could easily have been a Liar without being discarded by the subject, it is not knowledge

II SAFETY MEASURES

Since standard safety is normally advanced as a necessary condition for knowledge, these cases are not strictly counterexamples The point is that it arbitrarily focuses on the threat of error about *p*, when threats of error involving other propositions, as well as threats of emptiness and paradox, can undermine one's knowledge that *p*

Let us call any thought a *failed thought* if it has no content, or has as its content a false, gappy, or paradoxical proposition It is tempting to reformulate the safety condition by simply requiring that the very thought token whereby *S* believes that *p* could not easily have been a failed thought [11] But this hangs too much on the individuation conditions of thoughts, for instance, we would require the contentious assumption that contentful thought tokens do not have their contents essentially Without preempting these metaphysical issues, how can we clarify the idea of the near danger of "messing up?" We need a counterpart relation on thought, likely broader than identity, such that *S* knows that *p* only if

REVISED SAFETY *S* could not easily have had a failed counterpart thought

The idea is that one is safer the more distant are the closest worlds in which one has a failed counterpart thought Revised safety sets a threshold of closeness such that no belief can be knowledge if there are false or empty counterpart thoughts in worlds at least that close For those who are drawn to contextualism in epistemology, it is natural to hold that the threshold degree is context sensitive [12]

Note that we need a threshold for closeness of worlds *as well as* a counterpart relation for thoughts Very different untrue thoughts at

[10] The example is from John Hawthorne, "Deeply Contingent A Priori Knowledge," *Philosophy and Phenomenological Research*, LXV (2002) 247–69

[11] See the notion of "belief-safety" in Weatherson, *op cit*

[12] For gradable versions of standard safety with a contextually specified threshold, see DeRose, "Sosa, Safety, Sensitivity, and Skeptical Hypotheses," in John Greco, ed , *Sosa and His Critics* (Malden, MA Blackwell, 2004), pp 22–41, Sosa, "Relevant Alternatives, Contextualism Included", and Mark Heller, "The Proper Role for Contextualism in an Anti-Luck Epistemology," *Philosophical Perspectives*, XIII (1999) 115–29 If closeness only involves a partial ordering of worlds, the comparative will sometimes lapse

very close worlds do not count against knowledge, and neither do very similar untrue thoughts in very distant worlds We could gerrymander a similarity* relation for thoughts that builds in closeness of worlds, so that two duplicate thoughts in different worlds could differ in their similarity* to a third thought But the resulting concision is artificial, concealing as it does two quite different elements contributing to the notion of safety one having to do with similarity between thoughts, the other with overall similarity of worlds [13]

I will not attempt a precise analysis of the needed counterpart relation for thoughts, indeed, the vagueness of 'knows that p' may not allow precision in this endeavor But there are some clear structural constraints on the relation First, if empty thoughts can be counterparts of contentful thoughts, we cannot characterize the relation in terms of similar content Second, there is pressure to deny that two beliefs are counterparts if they are based on very different evidence Consider the following problem for standard safety discussed by Ernest Sosa [14] Suppose I see a large fish surface I recognize the distinctive dorsal fin of a sailfish, and I infer that a fish has surfaced nearby But a whale almost surfaced instead, and had the whale surfaced I would have thought falsely that a fish had surfaced In this case, while my actual belief fails the standard safety test, it seems to be a piece of knowledge—as long as I never mess up when confronted with sailfish-evidence

To handle this problem, Sosa suggests a version of safety along these lines

BASIS S could not easily have believed p on the same basis without p being true [15]

But this condition, like standard safety, does not rule out the threats at issue in section I Since the proposition I express by 'That is a lark' could not have been believed on any basis without its being true,

[13] For the latter I have in mind the metaphysical notion of closeness associated with, for example, the treatment of counterfactuals in David Lewis, *Counterfactuals* (New York Blackwell, 1973)

[14] See Sosa, "Replies," in Greco, ed , *Sosa and His Critics*, pp 275–325

[15] Sosa has implemented this general idea in different ways see Sosa, "Skepticism and Contextualism," *Philosophical Issues*, x (2000) 1–18, Sosa, "Replies," in Greco, ed , *Sosa and His Critics*; and Sosa, "Replies," *Philosophical Issues*, x (2000) 38–42 I have here interpreted his subjunctive conditional (the reason/basis for p would be present only if p) according to his suggestion in Appendix IV of "Relevant Alternatives, Contextualism Included" "One might hold that in a case where r is false, what is required for the truth of '$r \rightarrow p$' interpreted thus is not just that it be true that p in the closest worlds in which it is true that r, but that it be true that p in those worlds and also in other more remote worlds up to a certain (context-determined) threshold "

BASIS is satisfied in both the nearby-imposter and nearby-illusion cases But we can co-opt the idea behind BASIS by adopting REVISED SAFETY but requiring that counterpart thoughts must have the same—or a sufficiently similar—basis If we do this, we inherit a version of the dreaded generality problem for method-reliability conditions What counts as the same basis? Or how similar must the basis be? It would do no good to hold fixed, as Lewis does, every aspect of the subject's "evidential state " Suppose that in the first lark case, any of the imposters would have sounded ever-so-slightly different from any lark Nevertheless, I am unsafe—and not because of distant, skeptical worlds in which I hear exactly this sound but there is no lark [16] So the counterpart relation must cover beliefs with slightly different evidential grounds

It must also cover beliefs with slightly different content Consider Kripke's case of a region where all the fake barns are yellow, while the one true barn is red I look out the window, see the red barn, and come to believe that I am driving by a red barn [17] According standard safety, this belief is safe, while the (entailed) belief that I am driving by a barn is unsafe But intuitively my nearby possible false thoughts about yellow barns endanger my actual true thoughts about the red barn So these are similar enough to be counterparts [18] But a nearby false thought about how many sheep are on the hillside is not

It has often been thought that we need both a safety and a reliability condition on knowledge Reliability alone does not handle the

[16] It is worth noting a further difficulty for Lewis's account His claim is *S* knows that *p* iff *S*'s evidence eliminates all not-*p* worlds that are not being properly ignored ("Elusive Knowledge," *Australasian Journal of Philosophy*, LXXIV (1996) 549–67) The notion of evidence here is not understood in terms of what is known by *S* instead, we hold fixed all of *S*'s experiences, narrowly individuated But if we individuate the experiences narrowly and the proposition *p* widely, there is a problem Suppose *S* is looking at an enormous array of pixels, all dark except one ('alpha'), which is bright yellow *S* looks at alpha and says, 'That is lit' Moreover, there is a close, salient, experientially indistinguishable world in which the next pixel over (beta) was lit instead In that world, *p* is false (because pixel alpha is not lit) Of course, in that world *S* does not even believe that *p*; she believes that pixel beta is lit Nevertheless, Lewis's version of safety counts *S* as not knowing that *p* This is not a problem for Lewis if he intended that *p* be individuated narrowly, but it is a problem for those of us who take the semantic value of such demonstrative thoughts to be object-dependent See the related point and excellent discussion in Jonathan Schaffer, "Knowledge, Relevant Alternatives, and Missed Clues," *Analysis*, LXI (2001) 202–08

[17] Kripke's remarks from which this example is drawn remain unpublished

[18] Sosa suggests that one's belief that there is a red barn nearby is in some sense *based* on the unsafe belief that there is a barn nearby, whereas in the sailfish case the belief that there is a sailfish nearby is more direct (Sosa, "Replies," in Greco, ed , *Sosa and His Critics*) So in handling Kripke's case it is tempting for a proponent of safety to revise the account so that the necessary condition on knowledge is safety*, where a belief is safe* only if it and any transparent entailments are safe But the work can also be done by a notion of sufficient similarity of thoughts, which safety theorists must appeal to anyway

mitual lark case properly My ability to discriminate larks from other birds may be so reliable that there are only five token birds in the world that I would mistake for larks But if all five of them happen to be in my yard along with a real lark, that is enough to undermine my knowledge It is the *near* danger of messing up, not the unreliability of my faculty, that is at work in this case [19] At the same time, standard safety alone does not rule out coming to know a mathematical truth by tossing a coin with the rule heads, I will accept the Reimann hypothesis, tails, I will deny it [20] If this method generates a truth, that truth will be necessary, and so there will be no worlds in which I believe that very proposition and it is false Revised safety, in contrast, can treat nearby false beliefs about different mathematical propositions as counterpart thoughts It inherits a version of the generality problem, but unlike standard safety, it need not be supplemented by a reliability constraint

III REVISED SAFETY AND APRIORITY

Many philosophers are comfortable with fallibilism even when it comes to the phenomenology that typically accompanies a priori knowledge a thought that to all appearances is a piece of a priori knowledge could turn out to be false This can occur, for instance, if we are unlucky enough to have flawed internal mechanisms whose deliverances are accompanied with a strong sense of a priori obviousness But revised safety puts further strain on traditional notions of apriority for beliefs that would otherwise be paradigmatic pieces of a priori knowledge can be unsafe even when they are true and everything is going well internally

[19] To insist that my reliable bird-discrimination method becomes unreliable "in this environment" would be to operate with a notion of reliability-in-an-environment that requires infallibility

[20] DeRose in "Sosa, Safety, Sensitivity, and Skeptical Hypotheses," tries to motivate adding his "strength" condition with this sort of example Strength is disturbed by nearby worlds in which *S* disbelieves *p* but it is true But strength cannot handle a case in which I flip a coin to decide whether to judge the hypothesis true—or else to withhold judgment about it In that case I do not, in any close worlds, disbelieve the hypothesis One might be tempted to add "counter-safety"—in no close *p* world do I fail to believe *p*—but this is too strong In many cases of knowledge that *p* one might easily have been distracted and failed to believe *p* (see Sosa, "Replies," in *Philosophical Issues*)

In the belief-withholding coin-toss case, the problem is not that one might have been led to falsehood by the coin-tossing method itself The problem is that one might have been led to falsehood by whatever procedure caused us to adopt that method It is the meta-method, if you like, that is the problem (Suppose the subject is told, by someone she knows to be a great mathematician, that she cannot go wrong using this method, in that case, it may well deliver knowledge) And of course meta-methods, being themselves methods, can be handled by reliability or revised safety, so DeRose's extra condition is unmotivated

Consider an example due to John Hawthorne [21] "A Priori Gas" makes one liable to miscalculate without loss of confidence in one's conclusions (One becomes disposed to fail to "carry the one," perhaps) Now suppose I am walking through a dark alley, much of which is filled with A Priori Gas, but I do not actually enter the gas Had I passed through it at any point, I would have become temporarily disposed to miscalculate, without noticing the change As things stand, I am lucky enough to avoid it and I have no reason to suspect I am in the vicinity of A Priori Gas Now suppose that I calculate that 28 + 47 = 75 In a nearby world, I am in a cloud of gas and calculate 28 + 47 = 65, and if we allow that this is a counterpart thought, then by revised safety my actual true belief is unsafe Therefore I do not know that 28 + 47 = 75, even though everything has gone well internally [22]

More generally, consider any reliable method that allows us to know things a priori without inference, perhaps by causing us to find propositions primitively compelling Insofar as a reliable method is fallible, the gas associated with that method can exploit whatever it is that explains those rare occurrences in which the method misfires It follows that anyone who knows anything by that method will have a possible intrinsic duplicate with the same belief who is near the gas and so unsafe But these considerations should not lead us to skepticism about the a priori What matters for knowledge—even a priori knowledge—is whether the subject is actually safe, not whether being internally like the subject guarantees safety

The social nature of meaning leads to similar results for other paradigmatic pieces of a priori knowledge I have in mind a semantic picture according to which we all intend to use (even in our thoughts) public terms whose semantic value is settled by community-wide dispositions and patterns of linguistic use For example, take Jones's belief that every bachelor is eligible She believes this because she just finds it primitively compelling that bachelors must be eligible (If challenged, she would say eligibility is just part of what it is to be a bachelor!) And we can assume that she is right the Pope is not a bachelor But suppose that small changes in dispositions and patterns

[21] In conversation, much of the rest of this section grew out of conversation with Hawthorne about issues related to *Something in Mind*, *op cit*

[22] Objection 1 A Priori Gas temporarily alters cognitive structure, and close possibilities involving intrinsic changes in the subject should not matter to safety Reply This cannot be right in general, lest it make all our beliefs about our own intrinsic states trivially safe Objection 2 My "method" changes when I pass through a cloud, and thoughts derived from a different method should not be considered counterparts Reply If there are "methods" like addition-while-not-failing-to-carry-the-one, what keeps us from ruling out every unsuccessful calculation as the result of a different method?

of use in her community would have lead to a slightly different property being expressed by 'bachelor' Not everyone finds it primitively compelling to the same degree that bachelors must be eligible, and if a few more people had been disposed to think that noneligibles could be bachelors, the semantic value of 'bachelor' would have been a different property, bachelorhood*, which is compatible with noneligibility [23] Now, if Jones would still confidently have come to believe the proposition she would have expressed by 'All bachelors are eligible', she would have been wrong As a result, she is not safe in her actual belief that all bachelors are eligible [24]

Some will complain that insofar as Jones would not have been a normal user, she would not have counted as really knowing the meaning of the term 'bachelor' But in the envisaged example, there are no "normal users" of the term, just a range of intuitions about the whether, for someone to count as a bachelor, he must be eligible Moreover, users with different intuitions may be perfectly interspersed, so that talk of dialects is inapplicable [25] A further objection the example, as presented, assumes something like an epistemicist theory of vagueness Adopting an alternative picture instead would allow us to insist that if 'all bachelors are eligible' is true, very small changes in use of the sort relevant to safety could not make this sentence false—they could only induce vagueness in the predicate 'bachelor' and thus indeterminacy or lack of truth value in the sentence [26] There is a decision point here if we count a belief with indeterminate truth value as a failed belief, then Jones does not satisfy revised safety If not, it will be indeterminate whether Jones knows, because it will be indeterminate whether she is in danger of a false belief (In the latter case, the force of the example is somewhat mitigated)

In short, given safety and semantic externalism, it is possible to grasp the necessary connection between bachelorhood and eligibility and yet fail to know that all bachelors are eligible (Or at best—it may be indeterminate whether one knows it) Adopting parts of a public

[23] Ignorance due to semantic blindness of this sort is discussed in Williamson, *Vagueness*

[24] Of course, she is still safe in her belief that all bachelors are male

[25] This sort of example may lead people to become skeptical about semantics for public language But it is far from clear that limiting the semantic supervenience base to the use and dispositions of an individual does a better job of avoiding the mess Given our own indecision and inconsistency, not to mention our apparent reliance on semantic division of labor, the situation does not improve as the base is restricted

[26] One might say this if one were a supervaluationist, for example I cannot here consider how the many approaches to vagueness might handle this case But note the specter of higher-order vagueness, if small enough changes make a sentence indeterminate in truth value, there is pressure to concede that it is actually indeterminate whether the sentence is determinately true

language into one's mental life is a risky business, because introspection is not a perfect guide to the dispositions and intuitions that constitute the supervenience base for their meaning

Before considering a final example, let us distinguish two kinds of semantic externalism A linguistic item or thought is semantically external simpliciter just in case its semantic value does not supervene on the internal features of the speaker (thinker) But an expression or thought is *boldly* external just in case it might lack semantic value altogether in a sufficiently unfriendly environment [27] In the latter case, whether it has any semantic value, and not just which semantic value it has, depends in part on the subject's environment Let us assume that names introduced nondescriptively, and used in an ordinary way, are boldly external both in language and in thought As it happens, names almost always have semantic values in our linguistic community If I hear friends using the name 'Jane' in conversation, I might come to believe that 'Jane' refers to Jane or even that Jane is Jane, and it is rare for there to be an imminent danger of reference failure Indeed it seems that following the schematic rule

EMPLOY If 'N' is a name in your language and you have no reason to doubt that 'N' refers, infer that N is N

is (for me) a good way to arrive at knowledge But it would not serve me nearly so well were I unwittingly embedded in a language community with an alarmingly high proportion of empty names In that case, I might truly but unsafely believe that Jane is Jane or that 'Jane' refers to Jane For given bold externalism about these thoughts, they would have no content (or gappy content) in cases of reference failure It follows that I actually know, rather than truly believe, that Jane is Jane in part because of my friendly linguistic environment (Things do not go better if we fall back to 'If Jane exists, then Jane is Jane', given bold semantic externalism, this will also express a gappy proposition at best)[28]

However, the deleterious effects of an unfriendly environment can be counteracted Suppose that in the case where I pick up 'Jane' from

[27] I am pretending that these items are not individuated by content More carefully, a linguistic item or thought is boldly external iff a subject tokening such an item has possible intrinsic duplicates that token contentless counterparts of that item

[28] Granted, one approach treats all atomic gappy propositions as false, and would thus treat the proposition expressed by this conditional as true (As we have seen, it will also turn out to be the same proposition as that expressed by 'If Vulcan exists, then Bigfoot is Santa Claus') But aside from other difficulties with this approach, we have already argued in section I that even if some thoughts with gappy content are considered true, they should nevertheless be considered epistemic failures from the safety-theoretic point of view See footnotes 6 and 7 and the text to which they are appended

my friends, they had been plotting to fool me with an empty name but then ended up actually talking about a real person by that name In such a case I am only actually unsafe in my belief that Jane is Jane if I would have been fooled If certain cues would have made me suspicious enough to avoid adopting the name, then my actual belief is safe In such a case, let us say that I am *responsive* to the semantic dangers of my environment Note that one can benefit from one's responsiveness, or from a friendly environment, without realizing it Suppose I am asked, 'How do you know that *t* is not empty?' Having considered these matters, I may say 'I have empirical reasons to think that my community is not very unfriendly and also that if '*t*' had been empty, I would have noticed' If I am worried about the possibility of reference failure, I use this empirical evidence to rule out close content-failure scenarios even before being challenged

But consider Smith, who unlike me follows EMPLOY without relying on premises involving his friendly environment or his responsiveness (We will address the issue of *implicit* reliance in a moment) When he hears Jane being discussed among friends, Smith feels no obstacle in coming to believe such benign-seeming things as that Jane is Jane He simply has not reflected on the possibility of content failure This does not mean that Smith is unsafe indeed, we may suppose that his community is friendly and that he is (unselfconsciously) responsive to most of the danger that remains Likewise, we may suppose that Smith finds it primitively compelling that all bachelors must be eligible, and unlike Jones he is not in danger of semantic shift Can Smith reap the epistemic benefits of exploiting an external source of reliability without involving it in (even the tacit) justification of his beliefs? Can he enjoy this knowledge, blissfully ignorant of his semantic good fortune? As far as revised safety is concerned, yes, for to be safe does not require *knowing* that one is safe [29]

If we grant that Smith knows, we face a second question Should we count his knowledge as a priori? It might be argued that a priori knowledge cannot be environment-dependent in the way just described, but as we have seen, that would rule out even mathematical knowledge based on simple calculations To disallow Smith's knowledge as an instance of the a priori, it might be argued that unlike mathematical knowledge, it requires some sort of implicit justification by empirical premises Perhaps one cannot know that Jane

[29] But perhaps in order to *know* that one knows, one must know that one is safe In section IV, I discuss some of the implications of revised safety on iterated knowledge

is Jane or that bachelors are eligible unless one is somehow prepared to provide some a posteriori evidence to justify the belief if challenged with the possibility of reference failure or semantic shift, respectively

This raises delicate issues involving the nature of implicit justification For if someone challenges my belief that 28 + 47 = 75 on the grounds that I might be in A Priori Gas (or drunk) and therefore miscalculating without noticing it, I may trot out some empirical evidence that I am not in A Priori Gas (or drunk) Does this mean that I am always somehow implicitly relying on this evidence for my mathematical beliefs, or at least for my second-order belief that I *know* the mathematical truth?[30] One could try to sever the tie between answers to challenges and implicit justification, maintaining that Smith's beliefs are somehow implicitly justified by empirical evidence, while my mathematical belief is not—even though we both may appeal to such evidence when challenged But I doubt this distinction could be achieved with a natural conception of implicit justification For evidence about the reliability of our calculating skills is available to us all Indeed, it is difficult to imagine lacking such evidence, but if we did I suspect we would not be so confident in the deliverances of our calculations Does such evidence always play an implicit justificatory role in our beliefs? I have no good answer to this question My point in this section has been to illustrate with revised safety that these cases do not fit easily within the traditional distinction between a priori and a posteriori knowledge [31]

IV REVISED SAFETY AND SELF-KNOWLEDGE

Given revised safety and bold externalism for demonstrative thought, we can easily demonstrate that "strong privileged access" is false

(SPA) Necessarily, if S is thinking that p, then S is in a position to know a priori that she is thinking that p [32]

[30] Suppose it does One could argue that it is nevertheless *possible* for the mathematical belief (and even the second-order belief) to be justified without reliance on empirical data, whereas the belief that Jane is Jane could *never* be adequately justified without such reliance (except perhaps if the believer is Jane herself) Then, if one uses 'a priori' to mean 'not explicitly reliant on empirical justification, and in principle capable of being justified without even implicit reliance on empirical justification', one could insist that my mathematical beliefs are priori and Smith's beliefs about Jane are not

[31] I suspect it is vague whether the pieces of knowledge in question are a priori Terms of art are not after all immune to vagueness, especially those with so long and varied a history as 'a priori' I have no objection to putting 'a priori' to a precise use, as long as one acknowledges the technical nature of the term

[32] From Michael McKinsey, "Forms of Externalism and Privileged Access," *Philosophical Perspectives*, XVI (2002) 199–224 His version reads "x can in principle know a priori "

Happily, this can be done without relying on reductios of the sort that have been brought to light by Michael McKinsey and Paul Boghossian and have been discussed at great length in the literature on externalism and privileged access [33] And it can be done without worrying about whether introspective knowledge counts as a priori, because we can show that some thinkers that *p* are not in a position to know—a priori or otherwise—that they are thinking that *p*

Recall my demonstrative lark-thought in the case where I am actually hearing a lark but I have also been hallucinating lately We saw in section 1 that this belief is not safe Now suppose that by introspection, I come to form a thought I would express with 'I'm thinking that that is a lark' Note that in ascribing a demonstrative belief to me, this sentence itself *uses* and does not *mention* the demonstrative What would it express in the absence of a referent? If we hold, with semantic orthodoxy, that demonstrative expressions are boldly external even within the scope of propositional attitude ascriptions,[34] then this sentence would have failed to have any (complete) content And again, given bold externalism about the corresponding thought episode, it would likewise not have had any (complete) content It follows from this along with revised safety that if such a case is nearby, I do not know that I am thinking that that is a lark [35]

Unlike the standard arguments for the incompatibility of semantic externalism and SPA, the argument from safety does not concern the question whether it is possible to deduce facts about the world from a piece of introspection In fact, our subject does not go through any process of reasoning at all As a result, the most common ways to rebut McKinsey's and Boghossian's arguments have no application to the argument from revised safety

We have denied knowledge to subjects whose introspective beliefs are unsafe, but what should be said about those of us who practice

[33] See McKinsey, "Anti-Individualism and Privileged Access," *Analysis*, LI (1991) 9–16, and Boghossian, "What the Externalist Can Know A Priori," in Wright, Smith, and Macdonald, eds , *Knowing Our Own Minds* (New York Oxford, 1998), pp 271–84

[34] I am not assuming that propositional attitude contexts are "Shakespearian" namely, that co-referential singular terms can be substituted salva veritate in those contexts I am assuming that, whatever else this belief-ascription says, it must relate me to a singular proposition involving a specific lark or fail to have any complete content

[35] Brown considers the idea that the epistemic status of this sort of belief can be undermined by a close world in which I falsely believe something nonsingular, such as the proposition I would express by 'I am having a thought about some lark or other' See Brown, *Anti-Individualism and Knowledge* (Cambridge MIT, 2004) But given revised safety, the epistemic status of the singular introspective thought is directly undermined by the fact that in a close world, I token a thought internally just like it but lacking in content

safe introspection? Revised safety and object-dependence are compatible with a modified privileged access thesis, according to which if a subject knows that *p*, she is typically in a position to know that she knows that *p* So insofar as we take ourselves to know by introspection the content of many of our boldly external beliefs, we do face a self-knowledge puzzle after all

Here is a version of the problem using our terms [36] Suppose again that I have adopted the name 'Jane' after having heard much talk about a person by that name This time I am in a friendly environment, as well as tacitly responsive to the name-producing reliability of my community, like Smith in section III Then I can know by introspection

(1) I am thinking that Jane is great

Now, if I am aware that proper names are boldly external, and I know that I intend to use 'Jane' as a proper name whose referent is settled by my interlocutors, I am in a position to conclude that Jane's existence is a necessary condition for contentful Jane-thoughts

That is, I can know

(2) If I am thinking that Jane is great, then Jane exists

This coupled with knowledge of (1) appears to put me in a position to know that Jane exists Set aside whether we ought to count the premises as genuinely prior, it is puzzling enough that I am in a position conclude that Jane exists from an introspective premise along with a premise about a general feature of the semantics of my language

[36] The puzzle that follows is more similar to that of Boghossian, "What the Externalist Can Know A Priori," than it is to the one originally set forth in McKinsey, "Anti-Individualism and Privileged Access " McKinsey argues that from the truth of (a certain kind of) semantic externalism we can conclude that 'The proposition that Oscar is thinking that water is wet logically implies' a proposition involving the existence of objects external to Oscar (The emphasis here is on 'logically implies') But then, according to McKinsey, if Oscar can know a priori that he is thinking that water is wet, 'Oscar can just *deduce* E [the proposition about external objects] from something he knows a priori, and so he can know E itself a priori' In other words, Oscar himself need not know the truth of externalism, it just follows from externalism that he can deduce E from the introspective knowledge McKinsey reiterates the point in his "Transmission of Warrant and Closure of Apriority," in Susana Nuccetelli, ed , *New Essays on Semantic Externalism and Self-Knowledge* (Cambridge MIT, 2003), pp 97–116, on p 98

I find this version of the puzzle less compelling because it puts so much weight on a notion of "logical implication" between propositions (not sentence types), a relation that McKinsey says goes beyond metaphysical implication McKinsey requires that externalism commits us to the following thesis the proposition that Oscar is thinking that water is wet transparently implies some proposition about the existence of water (Where *p transparently implies q* just in case, if anyone knows that *p*, that person is in a position to simply deduce *q* from *p*) But this thesis needs more support than he provides

The puzzle should have a familiar ring to epistemologists it parallels Fred Dretske's classic closure puzzle Sandra, a normal zoo-goer, knows that the animal before her is a zebra So is she in a position to deduce that it is not a mule cleverly disguised as a zebra? (She knows the steps that would take her from one to the other) Or again, if you know that you will be going to work tomorrow, does this put you in a position to know that you will not win the "great surprise heart-attack lottery" tonight?[37] The self-knowledge puzzle is just a particular kind of closure puzzle in which the subject comes to know something that he should not be in a position to know by way of introspection plus general semantic knowledge A range of responses is available for any closure puzzle [38]

(a) Revoke the original knowledge The skeptical position
(b) Embrace the reductio The anti-skeptical conclusion
(c) Deny closure
(d) Adopt contextualism or "sensitive invariantism "

I will set aside (a) and (b) Skepticism is a last resort, and I do not want to allow that a good way of coming to know that I will not win the lottery (or die) tomorrow is by knowing that I am going on a modest vacation in a week That leaves (c) and (d)

In the case of the self-knowledge puzzle, we have two further options

(e) Deny bold semantic externalism
(f) Exploit gappy propositions to reject the move from (1) to (2)

Much of the literature on externalism and self-knowledge has focused on option (e), which as I will argue below is a red herring Moreover, while (e) has some plausibility when it comes to predicates like 'water',[39] which figured in the original McKinsey example, it is less

[37] For these and related cases, see Dretske, "Epistemic Operators," this JOURNAL, LXVII, 24 (December 24, 1970) 1007–23, Cohen, "How to Be a Fallibilist," *Philosophical Perspectives,* II (1988) 91–123, Jonathan Vogel, "Are There Counterexamples to the Closure Principle?" in Michael D Roth and Glenn Ross, eds , *Doubting Contemporary Perspectives on Skepticism* (Boston Kluwer, 1990), pp 13–27, DeRose, "Knowledge, Assertion, and Lotteries," *Australasian Journal of Philosophy,* LXXIV (1996) 568–80, Hawthorne, *Knowledge and Lotteries* (New York Oxford, 2004)

[38] Much of the recent literature on McKinsey's puzzle focuses on the issue of "warrant transmission " But note that the stance according to which the subject's warrant fails to transfer to the conclusion is not a further option to be added to (a)–(f)

[39] The idea would be that even if we have always been embedded in the worst sort of environment (such as Boghossian's "Dry Earth"), 'water' still has had some semantic value or other—for instance, the functional property *being clear, drinkable, liquid, and so on* See Boghossian, "What the Externalist Can Know A Priori " (Moreover, while having a belief about that property entails the *existence* of that property, it does not entail that

plausible for names and demonstratives A neglected alternative is (f) bold externalism only requires that 'Jane' would have no semantic value in the absence of a referent This is consistent with 'Jane exists' having gappy content We might then hold quite naturally that an attitude ascription of the form '*N* thinks that *S*', where '*S*' is replaced by a sentence with gappy content, is true just in case the individual referred to by '*N*' is belief-related to the gappy proposition expressed by '*S*' If this is the case, then knowledge of general semantic facts does not put one in a position to infer (2) from (1) [40]

However, I will set (e) and (f) aside because semantic externalism is something of a red herring in this vicinity To see this, consider that one can cook up closure puzzles involving knowledge of one's own mental states that have nothing to do with semantic externalism Consider my second order belief in the mathematical case

(3) I know that 28 + 47 = 75

If I know (3), and also that knowledge must be safe, this seems to put me in a position to deduce

(4) There is no A Priori Gas around [41]

This self-knowledge closure puzzle has just as much force as the original one, but (e) and (f) do not apply here Neither, of course, are they available for standard closure puzzles that do not involve self-knowledge It would be best to treat all closure puzzles in the same fashion, so that (to mix metaphors) we can kill many birds by biting one bullet And that is exactly what revised safety allows us to do, as I hope to show

Let us return to the option of denying

CLOSURE If *x* knows that *p* and *x* properly deduces *q* from *p*, then *x* knows that *q*

---

there are any *instances* of it) I will not contest this approach, except to point out that it seems much less plausible for proper names, demonstratives, and the like

[40] Things are more complex if in the example I know the second-order proposition expressed by 'I know that I am thinking that Jane is great' Then, if I know that 'Jane' is boldly external and that revised safety does not allow knowledge of gappy propositions, it would appear that I can infer (2) However, it might be suggested that even if 'Jane' is empty, the proposition expressed by 'I am thinking that Jane is great' should not count as gappy, at least for the purposes of revised safety For one could hold that the emptiness of 'Jane' does not hinder 'that Jane is great' from functioning as a complete term for a proposition, so in the relevant sense the belief ascription has a complete content

[41] Thanks to Hawthorne here

A few explanations have been offered for why closure might fail in these cases, but I will focus on the appeal to sensitivity as a necessary condition on knowledge

SENSITIVITY Had p been false, S would not have believed p [42]

If a piece of knowledge must be sensitive, it follows that even when a subject performs a proper deduction from a single known premise, the subject may not be in a position to know the conclusion For instance, Sandra's first belief about the zebra is sensitive, while the second is not

How can sensitivity help when it comes to the self-knowledge puzzle? The "bad" conclusion is that Jane exists But if Jane had not existed, I would not have believed that Jane existed For one thing, I would not have heard anyone using 'Jane' For another, given bold externalism, there would be no proposition about Jane to express, Jane being unavailable to serve as a component of it So there are no worlds at all, let alone a close world, where Jane does not exist but I believe that Jane exists For those worlds of evaluation *at* which the proposition is false are worlds *in* which it is not available to be the content of any thought [43] So sensitivity is no use

In fact, if sentences containing names and demonstratives typically have object-dependent contents, sensitivity cannot even do the work it was originally designed for Consider a slight modification of our original example with Sandra This time she uses a demonstrative in her thought she believes that *that* is a zebra and then concludes that *that* is not a cleverly disguised mule Assuming a plausible thesis of species-essentialism, both of these beliefs are trivially sensitive because they involve necessary propositions [44] This point is not a counterexample to sensitivity as a necessary condition on knowledge, but it is devastating for sensitivity nonetheless For the condition was *introduced* to block Sandra's inference from 'the animal is a zebra' to 'the animal is not a cleverly disguised mule' How can the condition be taken seriously if it does not also block the inference from 'that animal is a

[42] Proponents of sensitivity and its variants include Dretske, "Conclusive Reasons," *Australasian Journal of Philosophy*, XLIX (1971) 1–22, Robert Nozick, "Knowledge and Skepticism," in his *Philosophical Explanations* (Cambridge Harvard, 1981), pp 167–290, and DeRose, "Solving the Skeptical Problem "

[43] Alternatively, if we interpret the antecedent of SENSITIVITY as asking us to look at worlds *in* which p is false—rather than *evaluated at which* p is false—then the condition will be trivially satisfied because if p exists it cannot be false

[44] A proponent of sensitivity might suggest a theory of counterfactuals for which some (but not all) counterfactuals with impossible antecedents are false Even if such a theory could be made plausible and could give the right result in this case, it would only help with the demonstrative zebra case, not with the introspective closure case

zebra' to 'that animal is not a cleverly disguised mule'? Perhaps sensitivity can be revised (in the spirit of our revision of safety) to take care of these problems, but at the moment I do not see exactly how this could be done [45]

We are left with (d) The basic contextualist answer is straightforward Two speakers may utter a sentence of the form *S knows that p at t* about the very same *S*, *p*, and *t*, and yet one speaks the truth and the other speaks falsely In particular, while the epistemic credentials of a subject's belief might pass muster in one context, they might not be good enough when the possibility of error becomes salient So, while in ordinary contexts 'Sandra knows there is a zebra in the cage' is true, it is not true in a context where it is particularly salient that Sandra might have messed up In such a context, a different relation is expressed by 'knows', one that Sandra does not bear to *either* the proposition that there is a zebra in the cage, or the proposition that there is not a cleverly disguised mule in the cage The result is that, as Stewart Cohen puts it, "if we evaluate the closure principle relative to a fixed context, thereby fixing the standard, it comes out true "[46] So the letter of the closure principle, at least, can be observed

Unfortunately, contextualism is usually cashed out in a way that is not particularly amenable to the semantic picture we have been working with For suppose we flesh out the intuition about "salient chances of error" in terms of contextually-sensitive thresholds on the degree of safety required for knowledge, as Sosa, Keith DeRose, and Mark Heller do [47]

THRESHOLD For '*S* knows that *p*' to be true in a context *C*, *S*'s belief that *p* must be safe to a degree specified by *C*

Given a gradable version of standard safety, a belief is safer the more distant are the closest worlds where *S* falsely believes that *p* While this

[45] For other problems with sensitivity, see Sosa, "Skepticism and Contextualism", Vogel, "Tracking, Closure, and Inductive Knowledge," in Steven Luper-Foy, ed , *The Possibility of Knowledge* (Lanham, MD Rowman and Littlefield, 1987), pp 197–215, and Schiffer, "Contextualist Solutions to Scepticism," *Proceedings of the Aristotelian Society*, XCVI (1996) 317–33 Just to mention two (1) a belief of the form 'I am not wrong in thinking that *p*' cannot be sensitive, and (2) though my belief that I am not a brain in a vat is insensitive and therefore not knowledge (a result that sensitivity theorists like), my belief that I am not a brain in a vat with no auditory sensations is also insensitive and therefore not knowledge (a result that no one should like)

[46] Cohen, "Contextualism and Skepticism "

[47] DeRose, "Solving the Skeptical Problem," and "Sosa, Safety, Sensitivity, and Skeptical Hypotheses", Sosa, "Skepticism and Contextualism," and "How to Defeat Opposition to Moore," *Philosophical Perspectives*, XIII (1999) 141–53, Heller, "The Proper Role for Contextualism in an Anti-Luck Epistemology "

may get the right result for the standard zebra case, the explanation for why Sandra does not know the conclusion does not apply to the demonstrative zebra case, because in that case both beliefs will be maximally safe Neither can the explanation apply to the introspective closure puzzle my belief that Jane exists is also maximally safe, because in worlds where Jane does not exist I could not believe that she exists [48]

Here again revised safety is in its element

GRADABLE A belief is safer the more distant are the closest worlds in which a counterpart fails

Given a gradable version of revised safety, and applying THRESHOLD, we get the conclusion we want Take the demonstrative zebra case When ascribers are properly ignoring possibilities with cleverly disguised mules in them, they can truly utter, 'Sandra knows that that is a zebra' But the mere act of considering Sandra's chain of reasoning to the conclusion that *that* is not a cleverly disguised mule makes it impossible to ignore possibilities in which there is a cleverly disguised mule in front of her instead And in those worlds, Sandra has a false counterpart belief—in a different proposition The context has shifted, and now neither demonstrative proposition falls under the extension of 'knowledge' Closure, relativized to contexts, holds firm

Revised context-dependent safety also handles our self-knowledge closure puzzles In normal contexts, we may ascribe a priori knowledge to someone who thinks that Jane is Jane But when we consider the reasoning that leads to (2)—reasoning that involves semantic intuitions about the meaning of empty names and so on—we can no longer ignore the possibility that Jane has picked up an empty name This shifts our context and requires safety to extend out to such possibilities as well, which it does not So when we evaluate the status of the resulting belief that Jane *exists*, we are in a context where neither (1) or (2) count as known A similar story can clearly be told about my knowledge of (3) and (4) [49]

[48] The same thing is true of Lewis's approach, which tells us that *S* knows that *p* iff *S*'s evidence eliminates all not-*p* worlds that are not being properly ignored, and that different worlds are properly ignored in different contexts (Lewis, "Elusive Knowledge") Assuming object-dependence for demonstratives, this does not explain why we cannot count Sandra as knowing that *that* is not a cleverly disguised mule on the basis of her knowledge that *that* is a zebra For there are no not-*p* worlds to be ignored in either case

[49] For an application of contextualism to introspection puzzles, see Jakob Hohwy, "Privileged Self-Knowledge and Externalism A Contextualist Approach," *Pacific*

A related approach to closure puzzles is "sensitive invariantism"[50] According to this view, 'knows' does not express different relations in different speaker contexts, but whether S knows that p depends to a surprising degree on the S's own context For example, high stakes relevant to the belief and possibilities of error that are salient to the *subject,* can undermine knowledge This is not a variety of contextualism because 'knowledge' has but one semantic value in any context—it is just a relation that is at the mercy of practical features of the subject's environment The zebra closure puzzle is handled by saying that Sandra's knowledge of the original zebra-proposition dissipates as she starts to consider the possibility that the zebra is a cleverly disguised mule But closure need not be denied, as long as we are careful to understand it as the thesis that if S knows that p and S deduces q from p *while maintaining her knowledge of* p, then S knows that q

Once again the view needs to be tweaked in order to handle semantic externalism Hawthorne uses examples where subjects undermine their knowledge that p by considering indistinguishable possibilities in which p is false But we must be careful not to limit the sorts of nearby and salient possibilities that destroy knowledge to possibilities of *false belief that* p Otherwise, for reasons that should now be familiar, neither the demonstrative zebra case nor the self-knowledge puzzle could be properly resolved

Before concluding this discussion of self-knowledge, it is worth stressing that a subject can be safe in her self-knowledge even if she could easily have undergone an unnoticed semantic shift—as long as her thought would not have failed Safety thus does not require that, if S knows that S is thinking that p, then S must be in a position to discriminate the thought that p from any nearby thoughts with different content This is the right result, following Burge and others, I consider this sort of discrimination principle too strong[51] The ability to discriminate the proposition that one actually believes from other propositions is relevant to whether one counts as "knowing which"

Philosophical Quarterly, LXXXIII (2002) 235–52 Hohwy employs the contextualism of Lewis, "Elusive Knowledge," and so is subject to the objections of footnotes 16 and 48 above In particular, assuming bold externalism, his approach cannot handle any closure puzzles involving names and demonstratives, including the demonstrative zebra case and the case involving (1) and (2)

[50] See Hawthorne, *Knowledge and Lotteries*

[51] See footnote 17 above, see also Burge, "Individualism and Self-Knowledge," this JOURNAL, LXXXV, 11 (November 1988) 649–65, Kevin Falvey and Joseph Owens, "Externalism, Self-Knowledge, and Skepticism," *Philosophical Review,* CIII (1994) 107–37, Anthony Brueckner, "Ambiguity and Knowledge of Content," *Analysis,* LX (2000) 257–60

proposition one believes, but judgments of knowing which, like judgments of knowing who, are sensitive to an entirely different set of contextual parameters than are judgments of knowing that [52]

To sum up this section the self-knowledge puzzle should be treated as any other closure puzzle But sensitivity is of no use when it comes to closure puzzles of self-knowledge, and runs into trouble even with standard closure puzzles involving demonstratives Fortunately, pairing revised safety with either contextualism or sensitive invariantism allows us to resolve all of our closure puzzles in a single stroke

DAVID MANLEY

University of Southern California

[52] This contrast is worthy of more discussion than I can give it here But note that one can know a man but still not count (at least in certain contexts) as knowing who he is For example, you may get to know your next door neighbor quite well and not realize he is the son of Little Richard, in certain contexts where his paternity matters, you will not count as knowing who he is In analogous cases, one can know a proposition but still not count (in certain contexts) as knowing which proposition it is For more on the factors that govern the felicity of "knowing who/which" claims, see Steven E Boër and William G Lycan, *Knowing Who* (Cambridge MIT, 1986)

BOOK REVIEWS

All the Power in the World PETER UNGER New York Oxford University Press, 2006 xxix + 641 p Cloth $74 00

This huge book offers its readers a Humanly Realistic Philosophy, the central commitment of which is a quasi-Cartesian substantive dualism in which we human beings are souls that have free will We are empowered in the sense that we are part of a world in which there are real powers or propensities that are distinct from and irreducible to nonpower properties Hence, the powers that we have as souls—free will, for instance—are irreducible to the physical qualities we have of our bodies

The book's first half develops at length an ontology of real powers or Potencies, which are distinct from the so-called Qualities Unger challenges a view that he labels *Scientiphicalism,* which would be the claim that all the powers of things are derivative from their nonpower and spatial properties The second half of the book is concerned with substance dualism, the soul, God, the afterlife, and free will I did not understand, however, why these two halves had to be published as one big book Unger does not in my view adequately demonstrate the connection between the first and second halves I do, however, think that such a connection can be made It may or may not be the one that Unger had in mind

Despite the unconstrained size of the book, Unger fails to provide a convincing case for these various positions In this review, I attempt to discover why that is the case Before doing that, however, I feel I must comment on one issue straight away It has to be said that the writing style will not be to the taste of many readers, especially those who work in traditional analytic philosophy with its requirements of clarity and economy But having registered a stylistic distaste of the book, I will now try to set that aside and consider simply how plausible and interesting it is when considered as a piece of philosophy In that case I have to say that I think it is not very plausible and not nearly as interesting as Unger himself judges My diagnosis is that the book suffers from three methodological flaws Unger may well see these as virtues, as he frequently sets himself apart from what is produced by other philosophers But I think these are all issues where the standard analytic philosophy with which we are familiar has a distinct advantage

0022-362X/07/0408/424–31 © 2007 The Journal of Philosophy, Inc

The first flaw, which I take to be the biggest one of the book, is that persistently it is about what we (mainly *he*) can conceive or what possible story we can tell to explain a problem Conclusions are then drawn from what we can conceive to how the world is But this is to assume, which Unger does almost entirely without argument, that conceivability is a reliable guide to possibility The danger is that the book is about the workings of Unger's own mind, rather than the world itself One is left with the feeling that many of the worlds that Unger conceives are figments of his imagination This is no mere passing criticism, however The worlds Unger conceives are integral to the position advocated in the book, most importantly to the distinction Unger defends in the world between Qualities and Potencies

The scientiphical position he is criticizing has it that the powers or propensities of a thing are all physically derivative from the nonpower properties (8) In light of the book's second half, a special case of the scientiphical view that we should bear in mind is the case of the mental powers of persons, which would also be said to derive from the nonpower properties of their physical parts (9) Opposed to the view, Unger offers a richer metaphysic containing three kinds of basic property, irreducible to each other All properties belong to one and only one of these three categories (17) First, there are the spatial properties of things, the *Spatials*, consisting in sizes and shapes Second, there are the powers, for which Unger prefers the term *Propensity* (because *power*, I assume unacceptably, "connotes positive force" (13)) Third, there are *Qualities* or the phenomenal properties such as phenomenal blue or red It is characteristic of the scientiphical view to deny there are Qualities and one of Unger's first aims is to affirm this final category in his Humanly Realistic Philosophy It is at this point that our imagination and conceiving begins to do the work The basic idea is that we need Qualities to be able to form a conception of a physical thing (34) The methodological justification seems to be, therefore, that we humans should be at the center of an acceptable philosophy (that "man is the measure of all things" springs to mind) and chapter 2 lays the groundwork for this philosophy by gathering some basic human data, for example "I Am a Real Thinking Being and You Are Another" (36—Unger has a penchant for capitalization) Our conception of matter without Qualities is incomplete so no Humanly Realistic Philosophy should deny Quality The following "argument" is typical

> Anyhow, for us normally sighted humans, at least for folks whose conceptual powers are anything much like mine, we can't think of any spatial region, whether absolutely empty or whether fully material, without thinking of it as Colored somehow or other (81)

I really did not see why I needed colors to so conceive, indeed I did not understand why my conceiving had even to be on a visual model The notion of conceiving, indeed, seemed to play such a crucial role in the discussion yet was seriously underexplained with reliance resting on our pretheoretical understanding of it Unger's conceiving, nevertheless, seems the root of what I take to be the biggest metaphysical error of the book This is the contingency view in which

> With physical entities generally, there is *no necessary connection* between (1) precisely *which* Spatial Properties the thing has, and (2) exactly *which* Qualities the thing has, and (3) precisely *which* Propensities *comprise all* the Propensities of that physical thing (88)

This may well have been for some time an accepted, even orthodox view but such contingency is being increasingly challenged in the contemporary debate on powers Could the powers of a thing vary independently of its shape, for instance? Do the Propensities and Spatial properties of a thing float entirely freely of each other? Unger clearly thinks so but his argument is one I find wholly unconvincing What we are presented with is page after page (circa pp 88–308) of imagined worlds in which particulars have one set of Spatials and one set of Propensities and other worlds in which particulars with those same Spatials have a different set of Propensities While I would agree that there is some sense of "conceivable" in which these worlds are indeed conceivable, I do not see that these examples are any reliable guide at all to what is genuinely the case or genuinely possible Unger evidently thinks that they are, saying in one place "these allusions may be taken just as heuristically helpful performances, all in the service of articulating the variety that may obtain with physical individuals" (93) Unger is, therefore, inferring from what he can imagine to what is a real possibility (he considers how this can be justified only rarely, but see pp 132 and 180) But could a particular really be spherical without having the propensity to roll on a flat surface? One would think that this Spatial property and Propensity could not vary independently at all in spite of the manifold imagined other cases of independent variation Unger does produce an argument against the example I have cited Sphericality would not deliver the Propensity to roll because the spherical particular might not be rigid and might flatten out So the sphere rolls only if it also has the Propensity to retain its shape (269) and this Propensity cannot be derived from its sphericality But this is a bad argument If the particular changes from being spherical to being flat, then naturally it no longer has the power to roll But while ever it is spherical it will indeed have that aforementioned Propensity I am left feeling,

therefore, that "the variety that may obtain" could be just in Unger's imagination and not in reality What seems to count for Unger is that "we can tell a pretty good explanatory story" (606) but that still could leave this book's story a work of fiction

Perhaps this methodology has some kind of Kantian justification In placing humans at the center of his philosophy, his concern is with the physical as we must conceive it (555, for example) And "we humans can't do more than, or do better than, to conceive entities in time and in space" (618) Yet Kantianism is dropped in favor of a Spinozan view where reality is supposed to be far richer than we human thinkers can conceive (620) Thus, while we can understand that things can have extension in space and time, analogy suggests that there are many other, unknown dimensions (616)

What I suspect is at the root of Unger's dualist or trialist view is inadequate examination of the distinction between powers (dispositions, Propensities, capacities, and so on) and nonpowers (Spatials, categoricals, and so on) For such an overly-long book on the subject, it is suprising to find this key distinction so little characterized There are various views on offer in the contemporary literature on how we might draw a dispositional-categorical distinction One relevant view, to which Unger could have at least alluded, is the view that all properties are powers Those working in this field would call it the early-Shoemaker view If being spherical was just a cluster of powers, such as the power to roll, then there would be no way in which the spatial property and the power could vary independently The powers would be constitutive of the shape and thus essential to it Those who resist such a view will often do so on the grounds that the powers of a thing are a product not just of its shape properties but also of the laws of nature, which are contingent Unger certainly says things very close to this (95) But in the last decade or so such contingent laws have increasingly become an object of suspicion

This seems an appropriate point at which to move to the second main methodological flaw of the book It ignores, or refuses to engage with, the contemporary literature on the topic of powers Unger is open about this in the Preface and other places ("I'm not bothering to engage with very much of the contemporary literature" (107)) But he forces this home in a footnote to which I must object "And, please, don't get bogged down in almost all the rest of the recent literature" (108 n15) Now it is one thing to say this from a position of strength, knowing the literature and having fairly reached an assessment of its worth But Unger's book looks, in many places, as if he does not know the recent literature This he more or less admits in the Preface, citing as his main contemporary source on the subject

of powers some "phone conversations" with Charlie Martin While Martin is indeed a giant in this field, I do not think it is enough The problem then is that Unger labors over many problems that have been already considered elsewhere more clearly and succinctly For an author who likes to offer us so many options, shouldn't he at least mention other views of which he knows, even if just to discard them? Hence, we have an unnecessarily lengthy discussion of the "miracle" that many particulars of the same kind have the same Propensities It is indeed a puzzle on Unger's free-floating view of the relation between spatial properties and powers and he has to posit a completely fantastic hypothesis to explain it simply that the "proto-soup" before the Big Bang was propensitied to develop all these Propsensitively identical particulars (599 and 506)—a *virtus dormitiva* if ever I saw one At least some discussion of Brian Ellis's *Scientific Essentialism*[1] would have been useful here and could have prevented the drawn-out discussion If the powers of a thing are essential to its kind membership then the similar powers of all kind-members is no "miracle" at all Yet Ellis does not even appear in the bibliography (this bibliography is only around four pages long and one page-worth of those four are Unger's own works) Furthermore, how, in ignorance of the contemporary literature, can one claim one is developing "a philosophical position that's usefully very different from all already extant views"? (108) Rather, I would argue that until the last decade, the view that a thing's powers are distinct from and free-floating of its nonpowers was the most common view (see, for example, Elizabeth Prior's book *Dispositions*,[2] which also does not feature in Unger's bibliography) If anything, Unger's ontology leans on the side of being conservative so I really see no justification at all for Unger's imperative to his readers not to be "bogged down" in the recent literature, if we replaced the rhetorical "bogged down" with "know "

The second half of the book kicks off with the question of free will (chapter 6) where it is at first hard to see the connection between the book's first and second halves The ability freely to choose comes down, however, to a mental power that has no physical derivation (341) This is a power that is not subject to physical laws (356) Hence, the case for free will rests on the kind of separation between powers and spatial physical qualities that was discussed above If powers can float freely from other physical properties, then the power of free choice will not be constrained by brain states, for instance I have already indicated that I find this basic picture unappealing

[1] New York Cambridge, 2001
[2] Aberdeen University Press, 1985

It is in this second half of the book that the third major methodological failing comes to the fore Its lengthy discussions and digressions are almost always "settled" by Unger's unargued assertions or intuitions, and this is in my view a major reason why the book fails to persuade For example, on the issue of whether there can be experiences without an experiencer, we find that "there's no place for such absurd thoughts" (58) All too often, rhetoric is offered in place of argument On what looks like the conditional analysis of disposition ascriptions he offers us "this craving is a philosophically ill condition Get over it" (233) and says the position is "just so much philosophically superficial hogwash " The Problem of Evil is presented, though not under that name, as the reason not to believe in God, but it is again presented in rhetorical fashion Puppy dogs suffer, we are reminded So "A Benevolent Almighty Power, you used to say? Get real, buddy This is the planet Earth we're talking about, not a world that never was" (507) Another suggestion is dismissed as "utter hogwash" (501)

Of course, we have already encountered something similar with the first methodological flaw I identified Unger sets far too much stall by his own intuitions, hence within two pages he settles matters with phrases such as "At least to my mind" (twice) and "Heck, it sure seems so to me" (224, 225) Where there is a lack of arguments, there is little to argue against But Unger surprises us in chapter 7 by finally producing an argument It is not new but is his own old Problem of the Many The problem afflicts the physical but not the mental so there is a reason to support a form of Cartesian Dualism A cloud minus one water droplet is still a cloud, hence there are countless many clouds with overlapping locations just where we may have thought there was a single cloud (368–69) The same can be said of any physical particular But it cannot be said of an experiencer or a chooser We cannot say that there are countless many experiencers inside the one experiencer There is only one experiencer and, similarly, there is only one chooser At least this is an argument, then, but it is one that I cannot find compelling In the first place, the Problem of the Many is one we may want to resist (see David Lewis for a discussion[3]) But, in the second place, I was not convinced that if the argument applies to physical particulars, it should not also apply to mental particulars Couldn't there be billions of simultaneous experiencers (Unger calls them "schmexperiencers")? Unger says not but an actual argument is again hard to find (see, for example, p 402) Post-Hume,

[3] Lewis, "Many, But Almost One," in *Papers in Metaphysics and Epistemology* (New York Cambridge, 1999), pp 164–82

we have the notion of a mind as a bundle of experiences, in which case it would be not so dissimilar to the cloud As Unger puts it

> Crazy as it seems to me, quite a few current academic philosophers are quite prepared to live with the idea that, right now in my situation, there really are very many experiencers, with I myself being just one of the vast multitude (407)

But is this idea any more crazy than Unger's own view of physical objects? If they are a vast multitude, why can there not be very many minds within my body (or my many bodies)? Similarly, Unger raises and dismisses a Problem of the Many for free choice, asserting that there is only one chooser in me (it *is* me) But this seems based more on our "Humanly Realistic" intuitions than on any direct argument

We are, it emerges, souls, where each soul is equally and enormously well-endowed Propensitively (455) Some souls have better brains with which to work, however, explaining why some people seem brighter than others Our souls are generated by our bodies but can exist without them Unger protests that such a dualist view goes against the tide, with scientiphicalism being so popular, but the list of dualist views on p 461 shows that this position is actually quite fashionable At least Unger's own arguments are "entirely novel pieces of reasoning, quite unlike anything already available" (461), we are told It turns out that we may exist disembodied, though to no avail (chapter 8) because we would not be able to experience anything (496) Chapter 9 deals with the problem of how nonspatial souls exist when they are entirely inactive It emerges, after many pages of discussion, that Unger's answer to this problem is essentially the same as Descartes's We are always experiencing but sometimes are unconscious of those experiences (523)

The final chapter considers how souls are distinct Two physical particles that are qualitatively identical can nevertheless be conceived of as distinct because they occupy different locations of the spatial dimension But how could qualitatively identical souls be distinct if they have no spatial locations? Unger's answer is that, by analogy, there will be a spacelike dimension in which souls are located This dimension is called *sone*, and the sonal objects within it can be called *sonicles*, by analogy with particles (546) Despite many pages spent developing the analogy, however, it never becomes more than an analogy (he admits on p 559 that he has done little more than substitute one term for another and on p 616 that his argument is *only* analogical) so there is no compelling reason to believe that there is a sonal dimension What is more, the "solution" actually assumes an answer to the chapter's question, rather than demonstrates it ("we're

supposing each sonal soul to be sonally separate from every other soul" (560, my emphasis)) so I do not see how anything is achieved nor how the problem of distinguishing exactly similar souls is solved

At best, the view Unger presents of the world *could* be an explanation of it But offering an explanation of some phenomenon and arguing that it is the correct explanation are entirely different things The latter is in my view lacking from the book which consequently comes across as a flight of fancy In this review, I have tried to convey my overall impression that this book offers a very unfamiliar kind of philosophy I believe, indeed, that this has been Unger's intention But Unger also intends that his book be novel, interesting, and an overturning of the received view of professional philosophy It should be clear from this review that I have serious doubts that Unger has been successful in these respects

STEPHEN MUMFORD

University of Nottingham

NEW BOOKS

ACAMPORA, RALPH R *Corporal Compassion Animal Ethics and Philosophy of Body* Pittsburgh University Press, 2006 xv, 201 p Cloth $35 00

ADAMS, NICHOLAS *Habermas and Theology* New York Cambridge, 2006 ix, 267 p Cloth $75 00, paper $29 99

ADAMS, ROBERT MERRIHEW *A Theory of Virtue Excellence in Being for the Good* New York Oxford, 2006 ix, 249 p Paper $45 00

ALCOFF, LINDA MARTIN *Visible Identities Race, Gender, and the Self* New York Oxford, 2005 xviii, 326 p Paper $29 95

ALMOND, BRENDA *The Fragmenting Family* New York Oxford, 2006 359 p Cloth $24 95

AMBUEL, DAVID *Image and Paradigm in Plato's Sophist* Las Vegas Parmenides, 2007 xvii, 279 p Cloth $32 00

AMERIKS, KARL *Kant and the Historical Turn Philosophy as Critical Interpretation* New York Oxford, 2006 vi, 355 p Paper $39 95

ANDERSON, HANNE, PETER BARKER, and XIANG CHEN *The Cognitive Structure of Scientific Revolutions* New York Cambridge, 2006 xvii, 199 p Cloth $70 00

ANTONELLI, G ALDO *Grounded Consequence for Defeasible Logic* New York Cambridge, 2005 xii, 118 p Cloth $55 00

ARNOLD, DAN *Buddhists, Brahmins, and Belief Epistemology in South Asian Philosophy of Religion* New York Columbia, 2005 viii, 318 p Cloth $49 50

ARPLAY, NOMY *Merit, Meaning, and Human Bondage An Essay on Free Will* Princeton University Press, 2006 vii, 148 p Cloth $29 95

ASHLEY, BENEDICT M *The Way toward Wisdom An Interdisciplinary and Intercultural Introduction to Metaphysics* Notre Dame, IN University Press, 2006 xxi, 617 p Cloth $75 00

AUDARD, CATHERINE *John Rawls* Ithaca, NY McGill-Queens, 2007 viii, 328 p Paper $22 95

AUDI, ROBERT *Moral Value and Human Diversity* New York Oxford, 2007 xiv, 144 p Cloth $25 00, *Practical Reasoning and Ethical Decision, 2nd Edition* New York Routledge, 2006 xi, 249 p Cloth $105 00, paper $33 95

AUSTIN, SCOTT *Parmenides and the History of Dialectic Three Essays* Las Vegas Paramenides, 2007 xi, 98 p Cloth $28 00

AZZOUNI, JODY *Tracking Reason Proof, Consequence, and Truth* New York Oxford, 2006 vi, 248 p Cloth $49 95

BACHARACH, MICHAEL *Beyond Individual Choice Teams and Frames in Game Theory* Princeton University Press, 2006 xxiii, 214 p Cloth $35 00

BARNES, JONATHAN *Truth, etc* New York Oxford, 2007 x, 551 p Cloth $78 00

BAERGEN, RALPH *Historical Dictionary of Epistemology* Oxford, UK. Scarecrow, 2006 xlix, 255 p Cloth $75 00

BAGARIC, MIRKO *How to Live Being Happy and Dealing with Moral Dilemmas* Lanham, MD University Press of America, 2006 ix, 226 p Paper $29 95

BAGGER, MATTHEW *The Uses of Paradox Religion, Self-Transformation, and the Absurd* New York Columbia, 2007 xiii, 132 p Cloth $35 00

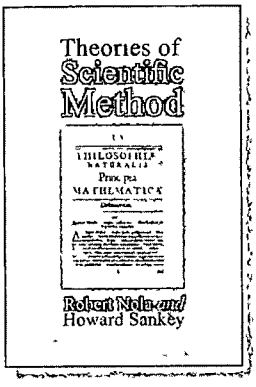
BAILEY, ALAN, and DAN O'BRIEN *Hume's Enquiry Concerning Human Understanding A Reader's Guide* New York Continuum, 2006 ix, 160 p Paper $14 95

MQUP Announces a New Series

Philosophy and Science

SERIES EDITOR ALEXANDER BIRD

This major new series in the philosophy of science aims to provide a new generation of textbooks for the subject The series not only offers fresh treatments of core topics in the theory and methodology of scientific knowledge, but also introduces newer areas of the discipline Furthermore, the series will cover topics in current science that raise significant foundational issues both for scientific theory and for philosophy more generally

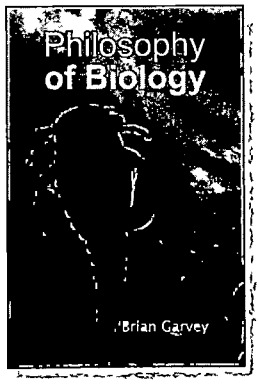


Theories of Scientific Method

Robert Nola and Howard Sankey

978-0-7735-3345-5 $27 95 paper | 978-0-7735-3344 8 $90 cloth

Informed by the latest thinking, students are offered an excellent introduction to the notion of scientific method and a wide-ranging discussion of how philosophers and scientists have grappled with the question over recent years



Philosophy of Biology

Brian Garvey

978-0-7735-3343-1 $22 95 paper | 978-0-7735-3342-4 $80 cloth

"An excellent introduction to the topic and deserves to be widely read Students will come away informed and excited about the subject "

–Michael Ruse, Lucyle T Werkmeister Professor of Philosophy, Director of the Program in the History and Philosophy of Science, Florida State University

MQ UP McGILL-QUEEN'S UNIVERSITY PRESS
www mqup ca

CAMBRIDGE

What's New in Philosophy

Race, Equality, and the Burdens of History
John Arthur
$80 00 Hb 978 0-521 87937 8 344 pp
$25 99 Pb 978 0 521-70495-3

Computability and Logic
Fifth Edition
George S Boolos, John P Burgess, and Richard C Jeffrey
$90 00 Hb 978 0 521 87752-7 368 pp
$29 99 Pb 978 0 521-70146 4

Music, Philosophy, and Modernity
Andrew Bowie
Modern European Philosophy
$99 00 Hb 978 0 521 87734-3 432 pp

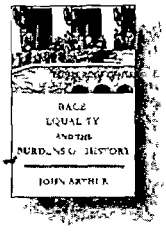
A Metaphysics for Scientific Realism
Knowing the Unobservable
Anjan Chakravartty
$85 00 Hb 978 0 521 87649 0 272 pp

Morality and Political Violence
C A J Coady
$80 00 Hb 978 0 521 56000 9 336 pp
$27 99 Pb 978 0 521 70548 6

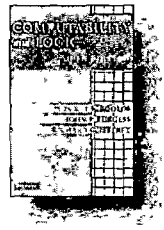
The Cambridge Companion to Carnap
Edited by Richard Creath and Michael Friedman and
Cambridge Companions to Philosophy
$85 00 Hb 978 0 521 84015 6 384 pp
$29 99 Pb 978 0 521 54945-5

Just Health
Meeting Health Needs Fairly
Norman Daniels
$80 00 Hb 978 0 521 87632 2 408 pp
$29 99 Pb 978 0 521 69998-3

Prices subject to change

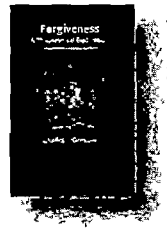


Logic Colloquium 2005
Costas Dimitracopoulos, Ludomir Newelski, and Dag Normann
Lecture Notes in Logic
$75 00 Hb 978 0 521 88425-9 288 pp



Pico della Mirandola
New Essays
Edited by M V Dougherty
$80 00 Hb 978 0 521 84736 0 240 pp

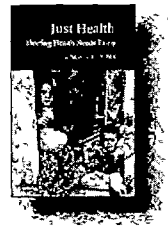
Common Law Theory
Douglas E Edlin
Cambridge Studies in Philosophy and Law
$80 00 Hb 978 0 521 84642 4 264 pp



Forgiveness
A Philosophical Exploration
Charles Griswold
$80 00 Hb 978 0 521 87882-1 272 pp
$21 99 Pb 978 0 521 70351-2

Hamann Writings on Philosophy and Language
Edited by Kenneth Haynes
Cambridge Texts in the History of Philosophy
$85 00 Hb 978-0 521 81741 7 292 pp
$24 99 Pb 978 0 521 52067-6

Socratic Epistemology
Explorations of Knowledge-Seeking by Questioning
Jaakko Hintikka
$80 00 Hb 978 0 521 85101-5 248 pp
$24 99 Pb 978 0 521 61651 5



The Cambridge Companion to the Philosophy of Biology
Edited by David L Hull and Michael Ruse
Cambridge Companions to Philosophy
$90 00 Hb 978-0 521 85128 2 552 pp
$34 99 Pb 978 0 521 61671-3

www cambridge org/us/philosophy

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE

from Cambridge

The Limits of Loyalty
Simon Keller
$85 00 Hb 978 0 521-87461 8 246 pp

Legal Ethics and Human Dignity
David Luban
Cambridge Studies in Philosophy and Law
$90 00 Hb 978-0 521-86285 1 250 pp

Plato on the Rhetoric of Philosophers and Sophists
Marina McCoy
$80 00 Hb 978 0 521 87863 0 224 pp

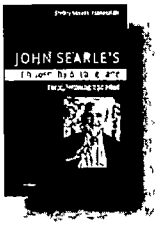
The Structural Evolution of Morality
J McKenzie Alexander
$95 00 Hb 978 0 521 87032 0 312 pp

Heine On the History of Religion and Philosophy in Germany
Edited by Terry Pinkard
Translated by Howard Pollack-Milgate
Cambridge Texts in the History of Philosophy
$85 00 Hb 978 0 521 86129 8 264 pp
$29 99 Pb 978 0 521 67850 6

Aristotle's *De Anima*
A Critical Commentary
Ronald Polansky
$100 00 Hb 978 0-521-86274-5 596 pp

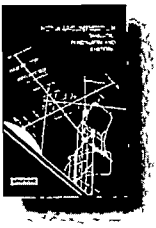
Analytic Philosophy and the Return of Hegelian Thought
Paul Redding
Modern European Philosophy
$95 00 Hb 978 0 521 87272 0 262 pp

The Cambridge Companion to Logical Empiricism
Edited by Alan Richardson and Thomas Uebel
Cambridge Companions to Philosophy
$90 00 Hb 978 0 521-79178-6 448 pp
$34 99 Pb 978 0 521-79628 6

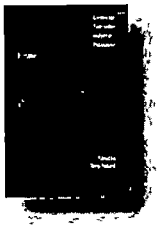


Bernard Williams
Edited by Alan Thomas
Contemporary Philosophy in Focus
$75 00 Hb 978 0 521-66216 1 240 pp
$24 99 Pb 978 0 521 66555-1

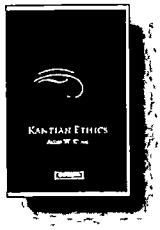
John Searle's Philosophy of Language
Force, Meaning and Mind
Edited by Savas L Tsohatzidis
$85 00 Hb 978 0 521-86627 9 312 pp
$34 99 Pb 978 0 521 68534-4



Media Argumentation
Dialect, Persuasion and Rhetoric
Douglas Walton
$80 00 Hb 978 0 521 87690 2 396 pp
$28 99 Pb 978 0-521 70030-6



Aristotle on Homonymy
Dialectic and Science
Julie K Ward
$85 00 Hb 978-0 521 87486 1 232 pp



Kantian Ethics
Allen Wood
$80 00 Hb 978 0 521-85494 8 360 pp
$25 99 Pb 978 0 521 67114 9

Prices subject to change

www cambridge org/us/philosophy

CAMBRIDGE UNIVERSITY PRESS

New from OXFORD

Why Think?

Evolution and the Rational Mind
RONALD DE SOUSA

"This book is a tour de force of scholarly insights on one of the most subtle puzzles in cognitive science—the relation between rationality and evolution"—Keith E Stanovich, author of *The Robot's Rebellion Finding Meaning in the Age of Darwin*

2007 160 pp , 2 line illus $25 00

New in Paperback!

Which Rights Should Be Universal?

WILLIAM TALBOTT

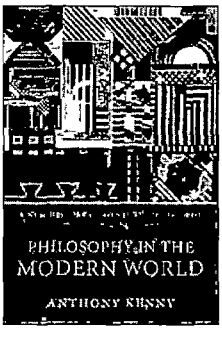
"No other work I am aware of comes close in making the consequentialist approach to rights come alive Talbott somehow manages to provide the most detailed and skillful account of the philosophical, institutional, and empirical complexity of this approach without ever letting us lose sight of the simple humanitarianism that motivates it"

— Liam Murphy, New York University

2007 256 pp paper $29 95 cloth $35 00

Philosophy in the Modern World

A New History of Western Philosophy, Volume 4
ANTHONY KENNY



Here is the concluding volume of Sir Anthony Kenny's monumental four-volume history of philosophy, the first major single-author narrative history to appear for several decades In this volume, Kenny tells the fascinating story of the development of philosophy in the modern world, from the early nineteenth century to the end of the millennium

2007 320 pp , 30 b/w illus $25 00

The Metaphysics Within Physics

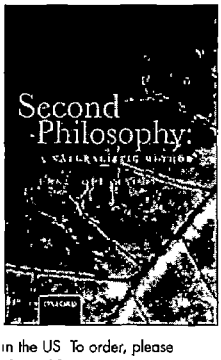
TIM MAUDLIN

Tim Maudlin's philosophical examination of the fundamental structure of the world as presented by physics challenges the most widely accepted philosophical accounts of laws of nature, universals, the direction of time and causation

2007 192 pp , 10 figures $49 95

Second Philosophy

A Naturalistic Method
PENELOPE MADDY



Many philosophers these days consider themselves naturalists, but it's doubtful any two of them intend the same position by the term Here Penelope Maddy proposes a particularly austere form of naturalism called "Second Philosophy"

2007 368 pp $65 00

OXFORD UNIVERSITY PRESS

Prices are subject to change and apply only in the US To order, please call 1 800-451 7556 In Canada, call 1 800-387-8020 Never miss an Oxford sale! Visit our web site at **www oup com/us**
Satisfaction Guaranteed or your money back

THE JOURNAL OF PHILOSOPHY

VOLUME CIV, NUMBER 9
SEPTEMBER 2007

page

433 *Compatibilism and Special Relativity* — Michael Levin

464 *Reasons as Causes in Bayesian Epistemology* — Clark Glymour and David Danks

COMMENTS AND CRITICISM

475 *Moral Responsibility without Alternative Possibilities?* — Carlos J Moya

487 NOTES AND NEWS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Wolfgang Mann, Mary Mothersill, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C Varzi *Consulting Editors* David Albert, James T Higginbotham, Robert May, Charles D Parsons, Wilfried Sieg *Managing Editor* John Smylie

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President*, Arthur C Danto, *Vice President*, Akeel Bilgrami, *Secretary*, Michael J Mooney, *Treasurer*, Liaquat Ahamed, *Other Trustees* Lee Bollinger, Leigh S Cauman, Barbara Gimbel, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

You may also visit our website at www journalofphilosophy org

THE JOURNAL OF PHILOSOPHY

2007

SUBSCRIPTIONS (12 issues)

Individuals	$45 00
Libraries and Institutions	$100 00
Libraries and Institutions (print and online)	$180 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for advertising rates, ad space is limited, so ad reservations are required

Published monthly as of January 1977, typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027 (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U S copyright law All rights reserved For more information about our copyright policies, please consult our January issue or contact us

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc , MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027 Periodicals postage paid at New York, NY, and other mailing offices

POSTMASTER Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027

© Copyright 2007 by the Journal of Philosophy, Inc

ISSN 0022–362X

THE JOURNAL OF PHILOSOPHY

VOLUME CIV, NO 9, SEPTEMBER 2007

ALMORE CALVIN

COMPATIBILISM AND SPECIAL RELATIVITY*

The future is now
—Attributed to George Allen, head coach of the Washington Redskins (c 1973)

It has lately been argued with new vigor that the Special Theory of Relativity (SR) rules out free will [1] The basis of this position, which I will call *relativistic fatalism*, is the framework-dependence of futurity In the space-time frame F whose origin O is my office at noon, July 1, 2007—which is now now—my grading in my office at noon while starting to want to run home in two hours precedes my running at 2 P M However, for some suitably chosen event E simultaneous in F with my grading there is a frame G in which E is simultaneous with my running Indeed—this was C W Rietdijk's version of the argument—for a suitably chosen E simultaneous in F with my grading there is a frame H in which E is simultaneous with an event that in F follows my running From my point of view at O in F I will run later, but in G I am running already and in H I have already

*I thank Margarita Levin for substantive and stylistic suggestions

[1] See Michael Lockwood, *The Labyrinth of Time* (New York Oxford, 2005), chapter 3, cited favorably by Galen Strawson, "The Bounds of Freedom," in Robert Kane, ed , *The Oxford Handbook of Free Will* (New York Oxford, 2002), pp 441–60, on p 441 The classic statement of the argument is C W Rietdijk, "A Rigorous Proof of Determinism Derived from the Special Theory of Relativity," *Philosophy of Science*, XXXIII, 4 (1966) 341–44, also see Hilary Putnam, "Time and Physical Geometry," this JOURNAL, LXIV, 8 (April 27, 1967) 240–47 Lockwood cites James Jeans and Einstein writing in the 1930s Palle Yourgrau also seems to favor relativistic fatalism (*The Disappearance of Time* (New York Cambridge, 1991), p 46), linking it especially to General Relativity with closed timelike curves, as does Lockwood later in *The Labyrinth of Time* (see pp 159–65, 327) I note that for Lockwood time loops involve fatalism "simply because of the truth of that tenseless view of time that time travel requires" (p 164) A corollary of the present paper is that compatibilism outflanks relativistic fatalism so long as causation is real, even if the possibility of time loops renders time "ideal" (as Yourgrau contends in *A World without Time* (New York Basic Books, 2005)), see footnote 27

0022-362X/07/0409/433–63 © 2007 The Journal of Philosophy, Inc

run So my future jog is as real and fixed as any event in my past, it is no more up to me whether to run at 2 P M than whether Socrates drank the hemlock Yet freedom demands a future that is open in a way in which the past and present are closed Hence I will (do) not run freely

Discussion of relativistic fatalism has tended to concentrate on ontological issues, such as the tense (if any) of statements locating events and entities in space time, the closure under simultaneity of 'Event E is real at worldpoint p', and the topology of elsewheres [2] Perhaps as a result, much of this literature presumes almost by default a libertarian conception of free will, according to which free agents must have access, in some suitably strong sense, to more than one possible future [3] SR is thought to undermine freedom because it is thought to undermine freedom of this sort Libertarianism is not the only idea of free will in play, however, there is also compatibilism Denying the libertarian presumption, compatibilists hold that freedom depends on how, not whether, behavior is caused, in particular on the causal role of preference Freedom—the real McCoy complete with "might have beens missed opportunities [and] avoidable mistakes"[4]—is therefore consistent with a causally fixed future For the same reason, a compatibilist may argue, freedom is also consistent with the relativistic fixity of the future The simultaneity of my running with an event simultaneous with the onset of my desire to run is irrelevant to my running freely, on this view, what matters, as always, is whether the cause of my running is a preference Since SR does not address the causes of human behavior, it cannot show that my running is not

[2] As well as the items in footnote 1, see Nicholas Maxwell, "Are Probabilism and Special Relativity Incompatible?," *Philosophy of Science*, LII, 1 (1985) 23–43, Rietdijk, "Special Relativity and Determinism," *Philosophy of Science*, XLIII, 4 (1976) 598–609, Steven D Hales and Timothy Johnson, "Endurantism, Perdurantism and Special Relativity," *Philosophical Quarterly*, LIII, 213 (2003) 524–39, Adolf Grünbaum, *Philosophical Problems of Space and Time* (New York Knopf, 1964), E J Lowe, "Time," in Ted Honderich, ed , *Oxford Companion to Philosophy* (New York Oxford, 1995), pp 875–76, Howard Stein, "On Relativity Theory and the Openness of the Future," *Philosophy of Science*, LVIII, 2 (1991) 147–67, Steven F Savitt, "There's No Time Like the Present (in Minkowski Spacetime)," *Philosophy of Science*, LXVII, Supp Vol , *Proceedings of the 1998 Biennial Meetings of the Philosophy of Science Association* (2000) S563–74, on S571 The issues of tense and simultaneity come up, obliquely, in section IV For our purposes a worldpoint may be an event, a construct from events, or an entity of some other sort Whatever their relation, a worldpoint and the events at it always receive the same coordinates

[3] A paradigm intrusion of libertarianism into ontology is Peter T Geach, "Some Problems about Time," in P F Strawson, *Studies in the Philosophy of Thought and Action* (New York Oxford, 1968), pp 175–91, see p 181 John Conway's "Free Will Theorem" also assumes libertarianism see http //homepage mac com/pmcarlton/iblog/C1074759898/ (March 21, 2005)

[4] Lockwood, p 69 Versions of compatibilism advanced by Harry Frankfurt, John Martin Fischer, Susan Wolf, and the Daniel Dennett of the 1980s deny or narrow our ability to do otherwise, in my view unnecessarily, see footnote 11 and section V below

caused by a preference, hence cannot show that I do not run freely The present paper develops this response to relativistic fatalism

Compatibilism has been said[5] to privilege everyday criteria for freedom while ignoring basic challenges to them Given how deeply SR challenges common sense, a charge of complacency[6] might seem especially apt against the response just outlined In the present context, however, such a charge puts the onus on the wrong party A logical link between SR and free will is prima facie doubtful, my freedom to run or continue grading appears in no way hostage to advances in fundamental physics The prior (second-order) probability that my freedom predicts observations inconsistent with SR, for instance, a phase shift in the Michelson-Morley experiment, is extremely low,[7] so low, in fact, that on a Bayesian analysis the denial of any such link enhances a theory's credence That compatibilism decouples free will from SR is a point in its favor

I

Crudely put, compatibilism holds that free actions are events caused by (the onset of) desire [8] More explicitly agent *S* does A freely[9] iff

[5] For instance, by Paul Russell, "Pessimists, Pollyannas, and the New Compatibilism," in Kane, ed , pp 229–56

[6] After writing this I found Saul Smilansky actually using "complacency"—"Compatibilism The Argument from Shallowness," *Philosophical Studies*, CXV, 3 (2003) 257–82, on p 271

[7] As low as that of a deterministic reduction of quantum mechanics expunging my guilt for my sins, for remarks in this same skeptical spirit, see Fischer, "Frankfurt-Style Compatibilism," in Sarah Buss and Lee Overton, eds , *Contours of Agency* (Cambridge MIT, 2002), pp 1–26, on pp 17–18, and Dennett, *Freedom Evolves* (New York Viking, 2003), pp 128ff Mark Balaguer asserts that "*neuro-cognitive science*" may show our choices are "'genuinely free'" in a "sort of libertarianesque way," on the grounds that "we are *colossally ignorant* on this topic"—"Libertarianism as a Scientifically Reputable View," *Philosophical Studies*, XCIII, 2 (1999) 189–211, on pp 196–97 He confesses he cannot imagine what that way would be like

[8] Onsets of desire, unlike states of desire, are events, so, if causes are necessarily events, onsets but not states of desire can cause action (Where possible I leave "onset of" tacit) Action may also be caused by the persistence or intensification of desire, or by the onset, persistence or intensification of other pro-attitudes, subtleties I do not try to do justice to Compatibilists should not say "choice" causes action For one thing, choices as traditionally conceived—spurts of pure will—are strangely elusive Avoiding choice also short-circuits the objection that nothing determined or believed to be determined can be chosen Finally, replacing choice by desire averts a familiar regress If the cause of a free action is itself free, as choices are said to be, the cause must have a free cause of its own A choice must be chosen, as must the choice of that choice Crude compatibilism stops this regress by flatly denying that desires are free Hierarchical compatibilism (see footnote 11) modifies this denial by counting desires caused by higher-order desires as free, but does not rescind it, since acts caused by undesired hierarchies of desire also count as free

[9] Departing, slightly, from the vernacular, compatibilism makes "do freely" a redundancy

A happens because *S* wants A to happen, *S* can do A iff A would happen if *S* wanted A to happen, and, if *S* does A, *S* could have done B ≠ A iff B would have happened had *S* wanted B to happen Here compatibilists understand 'X because Y' to require that Y be necessary as well as sufficient for X Idealizing somewhat, we may suppose that *S* never wants A to happen and also wants A not to happen Compatibilism then implies that *S* does A freely only if A would not have happened had *S* wanted A not to happen—that is, only if *S* could have done otherwise Compatibilists usually add that in fact people often do act freely and often can do otherwise, since much that occurs in the world in fact occurs because people want it to (and would not occur if they did not) Compatibilists are officially neutral as to whether all human behavior is causally determined, but it seems pointless for them to deny it [10] When a compatibilist dons his ever-handy determinist hat I will call him a *d-compatibilist*

Compatibilism faces well-known problems raised by compulsion, coercion, Frankfurt interveners, locked rooms, exogenous desires, deviant causal paths, finkish disabilities, and sour grapes, which, unaddressed, would leave its relation to SR moot Fortunately there are refinements of compatibilism (some also well known) that address these problems, but the details belong in a footnote [11] "The Minkowski

[10] Determinism is a major crossroads for incompatibilists Its acceptance leads to hard determinism, its denial and the assumption that there are free acts leads to libertarianism, its denial, denial that random acts can be free, and identification of indeterminacy with randomness leads to the a priori impossibility of freedom ("impossibilism") As a rule, all forms of incompatibilism converge in their criticisms of compatibilism

[11] (1) A compulsive gambler does not gamble freely despite wanting to gamble, because he impotently wants his desire to gamble to abate Hierarchical compatibilism therefore requires that a free agent's wants be polarized he must not mind his effective wants, not mind not minding, and so on (2) A victim of coercion does not freely comply with threats, despite both preferring and preferring to prefer compliance to costly defiance (Force of circumstance poses a similar puzzle) However, he typically minds not acting on his overall preference for cost-free defiance, an option removed by his coercer (or circumstance) Polarization can be stretched over these cases by requiring that a free agent not mind how his actions are explained Incidentally, compliance is an act of sorts, so not the paradigm of unfreedom it is often taken to be (by, for example, John Christman, "Introduction," in Christman, ed , *The Inner Citadel* (New York Oxford, 1989), pp 3–23, on p 9) (3) Suppose *S*'s endogenous taste for wine over beer preempts intervention that would otherwise have created this taste Compatibilism agrees with common sense that *S* drinks wine freely, but also insists, contrary to the received view, that *S* can do otherwise *S* can drink beer, since he would drink beer if he wanted to (a counterfactual that tracks back to intervener failure) The existence of the intervener destroys *S*'s power to *want* beer over wine, say compatibilists, for were *S* to try to acquire a taste for beer he would, thanks to the intervener, fail Compatibilism sharply distinguishes freedom to do from freedom to want (4) While often conflated (for example, by Fischer, "Frankfurt-Type Examples and Semi-Compatibilism," in Kane, ed , pp 281–308, see pp 282, 293, 295, and 304), inert intervener cases differ from cases in which the agent's desire is sufficient but unnecessary

view of space-time," as a challenge to "any conception of free will that pictures human agents as selectively conferring actuality [on potential futures]," as described by Michael Lockwood (*op cit*, p 69), is not apt to discriminate among conceptions of free will that picture agents otherwise It would be astonishing if crude compatibilism fell to the

---

for what ensues, as when *S* happily stays in a room he does not know is locked Here *S* truly can do no other than stay, yet appears to stay freely Locke says that he stays voluntarily but not freely (*Enquiry* II, XXI, 4), Frankfurt that he stays freely but is not free to go I suggest saying that *S* stays *weakly* freely Weak freedom is of interest because it reveals character, and thereby resolves the puzzle of moral appraisal in the absence of (full) freedom, for instance, chastisement of a feckless lifeguard for not trying to save a child who would have drowned before he reached her (5) According to compatibilists, Martians planting a polarized hierarchy of desires in *S*'s brain create a new, free, responsible person S^* Contrary intuitions are explained pragmatically S^* can be controlled by reinforcement, but the spread of the implanted desires to others is best controlled by reinforcing the Martians, not making an example of S^* Since imputation of responsibility implicates readiness to reinforce, and reinforcing S^* is relatively useless—it cannot deter or inspire others—we decline to call S^* responsible (6) Compatibilists (like others) require that the pathway from desire to action include feedback to assure adherence to a plan Elated by his resolve to bid on a painting, *S* punches the air, unbeknownst to him a bid under auction house rules He does not bid freely because he did not want to bid that way, and would have desisted had he noticed his gesture catching the auctioneer's eye (7) *S* is not trying to raise his paralyzed arm, but if he tried an angel would relax the paralysis and he would succeed Apparently, *S* cannot raise his arm but would raise it if he tried (See Keith Lehrer, "An Empirical Proof of Determinism," in Lehrer, ed , *Freedom and Determinism* (New York Random House, 1967), pp 175–202, Bernard Berofsky, "Ifs, Cans, and Free Will The Issues," in Kane, ed , pp 181–201, see pp 191–93) In fact, *S* can raise his arm or would fail if he tried Holding his paralysis constant while letting effort vary, his arm would stay put if he tried to raise it Letting paralysis vary with effort, the intent of the example, he *can* raise his arm—by inducing the angel to relax his paralysis by trying to raise it We confusedly think *S* cannot raise his arm but would do so if he tried because we waver between fixing and unfixing the paralysis, and we waver because *S*'s failure to try is so vividly presented as *sufficient* for paralysis that we half-forget that failure to try is also *necessary* (If not trying were unnecessary, the angel might disregard trying) So we half-forget that supposing *S* to try cancels anything for which not trying is necessary, such as paralysis, and half-think effort and paralysis are cotenable Compatibilism treats unconsciousness and phobias similarly· supposing a sleeping ailurophobe wants to pet a cat cancels his sleep and fear (8) Compatibilism implies that *S* becomes free by ceasing to want what he can not get and embracing ("identifying with") realizable wants he cannot curb Whatever the Stoics and Spinoza may have meant by similar-sounding doctrines often held up to ridicule, compatibilists are merely distinguishing what *S* is *free to do* at a time from whether he is then *acting freely* If *S* stops trying to do A, where A is something he cannot do, and starts trying to do something he can do, he starts acting freely This outcome is confusing because *S* remains unable to do A and refrains from A only weakly freely Commonly, one becomes free by gaining the ability to do what one persistently tries, *S* has become free without gaining any new power Even less common is attainment of freedom without gain in power *or* change in overt behavior, but that happens when *S* stops trying to resist an irresistible desire He then stops trying to do a covert thing he can not do, namely resist Compatibilists, by the way, are not so naive as to think that a desire can be quelled or embraced at will (Point 8 reflects discussion with Ellen Fridland and Richard Sorabji)

Minkowski view but a refined variant survived—if, say, actions done from undesired desires turn out to be unfree because these actions coexist with the desires that cause them, yet actions done from desired desires somehow turn out to be free despite coexisting with the (desired) desires that cause them Nor will relativistic fatalists wish to contest compatibilist treatments of these puzzle cases, lest their fatalism drift too far from its empirical mooring in SR So a focus on crude compatibilism misses nothing essential, and keeps the essentials in plain sight [12]

Lockwood might seem to have anticipated the compatibilist response, since he seems to concede the efficacy of choice [13]

> It is a truism that we ordinarily think of ourselves as causing things to happen, by way of our decisions and consequent actions As a matter of everyday experience, we find that, when we do something with the intention of satisfying some wish, the wish routinely (though by no means infallibly) comes true [W]e are surely still entitled, *pace* Jeans, to think of ourselves as shaping the future I need not deny, for example, that my action in hitting the golf ball with a Number 9 iron was caused by my choice so to do

However, as the oddly passive phrase "we find that [our wish] comes true" portends, his concession is hedged and equivocal He denies that "Nature falling into line with [our] wishes" is enough for freedom on the grounds that

> causing something to happen, in its ordinary sense, means so acting as to transform a potential occurrence into an actual occurrence For if the future course of events is already, in its entirety, an actuality, then nothing we ever did could possibly *count* as transforming a potential future occurrence into an actual one I can no longer think of myself, here, as genuinely adding items to the inventory of the real (*op cit*, p 55)

So Lockwood is not granting that choices may cause behavior, and then arguing from the actuality of the future that behavior caused by choice is nonetheless unfree He is arguing, from the actuality of the future plus an Aristotelian potency-into-act account of causation, that choices never cause behavior Ironically, rejection of compatibilism for this sweeping reason is more apparent than real It leads immediately to denial that there are any causes at all, transmuting the issue

[12] All forms of d-compatibilism face the consequence argument my running is not up to me if causally determined by laws and a past that are not up to me The argument is a non sequitur, since being up to me, when identified with being caused by my wants, is open under "is an effect of" Compatibilists take recent discussions of the argument to have made clear that determinism unambiguously entails "Nothing is up to me" only when "up to me" means "undetermined," reducing the entailment to the tautology that causally determined events are causally determined

[13] His word, see footnote 8

into—or revealing it to be—the existence, not of freedom, but of causality After all, the constancy of inventory is said to be global, and Lockwood himself is quite explicit that causation per se requires actualization of potentiality (hence an A-series) [14] If a choice C at t cannot produce an act A at $t^+ \geq t$ because A, being already actual at t, cannot be the one potentiality among others that C actualizes, then by parity of reasoning no event E at t can actualize any event E^+ at t^+, for E^+, also already real at t, is not a potentiality for E to actualize (Indeterministic causation must also go However many alternatives to E^+ might putatively have been actualized by E at t^+, E^+ cannot have been the one actualized if already actual at t) The conclusion is inescapable that nothing causes anything

Relativistic fatalists ready to bite this bullet might think twice before chomping down They are certainly not compelled to renounce causality by SR alone, which respects standard conditions on the time order of cause and effect (see section II) They may therefore find it less disruptive to adopt a non-Aristotelian account of causality consonant with SR, such as energy transmission, and confront compatibilism restated in its terms Moreover, denial of causality wars with their larger goal, shared by all fatalists, of showing that people are far more constrained than ordinarily supposed [15] If nothing is causally impossible—causality being an illusion—people are *less* constrained than ordinarily supposed From the fatalist's own viewpoint, Lockwood's argument proves too much

With the Aristotelian garnish removed, what remains of relativistic fatalism is the original meat-and-potatoes recipe no one has a say in what will happen because it is happening already But with SR again admitted to be consistent with causation, SR must again be admitted to be consistent with the causation of behavior by choice Lockwood must therefore still explain (as he has so far not even tried to do) why a future act caused by a present choice cannot already exist *and* be free, why the present existence of this act makes it any less *mine* [16] Perhaps a non-Aristotelian relativistic fatalist can offer an explana-

[14] "[S]uppose that determinism is true, so that at no time is there ever a possibility for anything to happen other than what eventually does happen Even so, the essence of this tensed account of time remains intact For there is a sense, clearly, in which we can regard every occurrence—even one that is inevitable—as the actualization of a potentiality The coming-to-pass of an inevitable event, we might say, is merely a limiting case of potentiality, where there is a priori potentiality for the event to happen and, by default, the potential for it to happen is in due course actualized" (p 18)

[15] "The driving force behind incompatibilism is the idea that the laws of nature constrain our abilities"—Helen Beebee, "Local Miracle Compatibilism," *Noûs*, XXXVII, 2 (2003) 258–77, on p 273

[16] Mark Zelcer put the issue this way

tion, sections II and IV allow him to develop his argument in detail But Lockwood has simply failed to engage compatibilism

It should be noted that Lockwood himself, like Rietdijk,[17] implicates acceptance of causation (perhaps inadvertently) He asserts that relativistic fatalism is consistent with indeterminism, an apt assertion only if relativistic fatalism does not entail the absence of causation If relativistic fatalism entails absence of causation, its consistency with the absence of causation goes without saying In the same connection we might observe that, if SR opposes free will along a broader front than does classical determinism, it should raise difficulties for compatibilism that determinism does not We will find this expectation disappointed

II

Compatibilists do not dispute the physics of relativistic fatalism Where $T(E, F)$ is the time of event E in frame F, it is a fact that, for any E_1, E_2 and F such that $T(E_1, F) < T(E_2, F)$, there is an event E_3 and a frame G such that $T(E_3, G) = T(E_1, G)$ and $T(E_3, G) \geq T(E_2, G)$ At issue is the significance of this fact, so a review of its derivation is in order

For convenience, set the speed of light at 1 Let the origin O of frame F coincide with the origin O′ of the rest frame F' of an electron moving with velocity v in F, so that in both frames the coordinates of the coincidence event are (0, 0) Experiment indicates that the unit time (the "hour") of F' as measured in F exceeds the hour of F by the factor $1/\sqrt{1-v^2}$ It follows that $\Delta t' = \Delta t\sqrt{1-v^2}$, where Δt is the interval between two events as measured in F, and $\Delta t'$ is the interval between the same events as measured in F' Experiment also indicates that the unit length (the lighthour) of F exceeds the lighthour of F' as measured in F by the factor $1/\sqrt{1-v^2}$, so that $\Delta x' = \Delta x/\sqrt{1-v^2}$, where Δx is the distance between two events as measured in F and $\Delta x'$ is the distance between the same events as measured in F' $\sqrt{(\Delta x)^2 - (\Delta t)^2}$, the spacetime separation of two events in F, is easily seen to equal $\sqrt{(\Delta x')^2 - (\Delta t')^2}$, their separation in F' Separation is invariant

[17] "[W]e may conclude that there is determinism (which is, of course, not the same thing as causality), also in microphysics Also there is no free-will, In the same way, of course, we can conclude that theories, intending to save free will with the aid of 'indeterminism' in micro-physics, cannot be tenable" (Rietdijk, "A Rigorous Proof," p 343) In fact, Lockwood endorses what I later call B-series fatalism "The tenseless view, after all, requires us, quite independently of relativity, to regard all events as being timelessly real, whether they are to be thought of as being laid out in time or in spacetime It may or may not be true, of course, that the future course of history is *determined* by what has already happened" (p 67, replying to an objection to relativistic fatalism) Joshua Dulberger pressed me on whether in the end relativistic fatalism differs from classical determinism

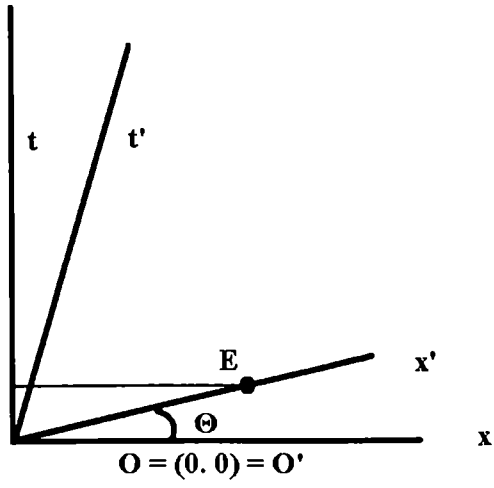


Figure 1 T(Coincidence, F) = T(Coincidence, F) = 0

These identities are packed into the Lorentz map, which assigns to a worldpoint with F-coordinates (x, t) the coordinates $((x-vt)/\sqrt{1-v^2}, (t-vx)/\sqrt{1-v^2})$ in F' If x and t are the space and time axes of F, the space and time axes of F' can be depicted as the lines x' and t' through O making angle $\theta = \arctan v$ with x and t respectively (In effect, x' is the hypotenuse of triangle xOx' and t is the hypotenuse of triangle tOt') The x' axis consists of those worldpoints whose time coordinate t' in F' is 0, that is, those worldpoints that in F' are simultaneous with O More generally, n is the time coordinate in F' of any point whose coordinates in F are $(x, vx + n\sqrt{1-v^2})$ When $v = 6$, for instance, an event E at (3 33, 2) in F is at (2 65, 0) in F', as in Figure 1 Thus, if F is my office frame, what occurs 2 hours after (and 3 33 lighthours from where) I grade in F occurs in F' when I grade

To complete the fatalist's argument, pick any event at $x = 0$ after $t = 0$ in F, such as my starting to run R, whose F coordinates are (0, 2) Next pick any worldpoint on the x axis (where $t = 0$) making an angle $\theta < 45°$ with R For some $v < 1$, $\theta = \arctan v$ and $|x| = t/\tan\theta$ Now let F' be the frame whose origin O' is moving with velocity v in F at the worldpoint whose coordinates in F are $(-t/\tan\theta, 0)$ In F', R is simultaneous with events whose F coordinates are $(-t/\tan\theta, 0)$ If V is the presence at (–3 33, 0) in F of an electron e_v moving at $v = 6$ in F, and P is the onset at (0, 0) in F of my desire to run, T(V, F) = T(P, F) and T(V, F') = T(R, F') V is simultaneous with P in F and with R in F', as in Figure 2

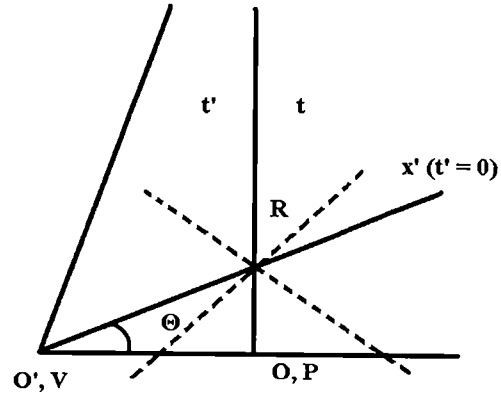


Figure 2 $v = dx'/dt = 6$

I will call PVR a *simultaneity fork* In general, a simultaneity fork $E_1E_2E_3$ is a trio of events E_1, E_2, E_3 and a pair of frames F, F' (often clear from context) in which $T(E_3, F) > T(E_1, F) = T(E_2, F) = T(E_2, F') = T(E_3, F')$ Now let $X(E, F)$ be the position of event E in frame F, a frame F is proper for a pair of events E, E′ when $X(E, F) = X(E', F)$, some representative such frame (for instance, the one in which $T(E, F) = 0$) may be called the proper frame $F(E, E')$ [18] The (absolutely) *early* tip of fork $E_1E_2E_3$ is the earlier of $\{E_1, E_3\}$ in $F(E_1, E_3)$, and the fork's *late* tip is the later of $\{E_1, E_3\}$ in $F(E_1, E_3)$ In particular, P and R are the early and late tips of PVR W, the presence of an electron e_w moving at $0 < w < 6$ in F at the worldpoint with F-coordinates $<x = -3\ 33,\ t = 0>$, is simultaneous with P in F and, in e_w's rest frame, simultaneous with an event Q that in F occurs between P and R For $w = 27$, Q, the late tip of fork PWQ[19] might be my questioning at $x = 0$, $t = 1$ whether to burn my running shoes to ensure my continued grading The presence of an electron e_z moving at $1 > z > 6$ at $<x = -3\ 33,\ t = 0>$ in F is simultaneous with P in F and, in e_z's rest frame, with an event S that in F postdates R, 9 forming a fork PZS

Compatibilists grant that every event in my proper future is the late tip of a simultaneity fork whose early tip is my present They therefore grant that my future already exists, so long as "my future

[18] There are no proper frames for spacelike connected events
[19] Although PWQ ≠ PVQ, $T(Q, F) = T(V, F)$ for all F

already exists" means precisely that every event in my proper future is the late tip of a fork whose early tip is my present They also grant, with the same caveat, that relativistic fatalism is at least initially clear where classical block universe fatalism is obscure Recall that the classical block universe is an eternal B-series of Newtonian moments—parallel sections of 3-space—without A-series becoming B-series fatalists maintain that an event E occurring at a time t already exists at all times $t^- < t$, hence that the occurrence of E at t is inevitable at all earlier times The obscurity is that, in B-theory, 'E at t exists at $t^- < t$' is ill-formed It cannot mean that E occurs (timelessly, eternally) at t^-, since E occurs at $t \neq t^-$ It cannot mean that E's occurrence at t occurs timelessly at t^-, for while events occur at times, as do occurrences of events (assuming the two distinct), the central tenet of B-theory is that the occurrence of an event at a time is timeless [20] Occurrences of events at times are not themselves events To be sure, 'E at t exists at t^-' can be taken to assert, coherently, that 'E occurs at t' is true at t^-, but the truth at t^- of 'E occurs at t' implies fatalism only if the corresponding A-series truth also implies fatalism For suppose, switching to A-theory, that it is now t^- and that E will occur, later, at t Then 'E will occur at t' is true now,[21] and if the truth at t^- of 'E occurs at t' necessitates E, so does the truth now of 'E will occur at t' Appeal to the B-series, when coherent, is thus superfluous Worse, as is widely (not universally) appreciated, these equipollent A- and B-truths do not in fact yield fatalism Let '$I_t(p)$' abbreviate 'p is inevitable at t' I_{now}(E will occur at t)[22] follows from the A-series truth 'E will occur at t' and the A-series truism

I_{now}('E will occur at t' is true $\supset$ E will occur at t)

[20] See Nathan Oaklander, "Freedom and the New Theory of Time," in Robin Le Poidevin, ed , *Questions of Time and Tense* (New York Oxford, 1996), pp 185–205

[21] Dummett might deny that 'E will happen' is true now even if E will happen He does deny that 'p' always entails '"p" is true', on the grounds that 'There was no King Arthur' may be the best response to 'King Arthur defeated the Saxons' but not to 'The statement "King Arthur defeated the Saxons" is true' ("Truth and the Past," this JOURNAL, C, 1 (January 2003) 5–53, on p 21) However, this disparity does not mean that the best response to 'The statement "King Arthur defeated the Saxons" is true' is its negation, that is, '"The statement 'King Arthur defeated the Saxons' is true" is false' A man who withholds a truth value from 'King Arthur defeated the Saxons' may doubt that it makes a statement, so may respond to 'The statement "King Arthur defeated the Saxons" is true' with 'There is no such statement' (Kreisel mimics this man, less a pair of quotation marks 'Is it true that the King of France is bald? There is no King of France, so what are you talking about?'—"Bertrand Russell's Logic," in D F Pears, ed , *Bertrand Russell* (New York Doubleday, 1972), pp 168–74, on p 171)

[22] The subscript 'now' is redundant, $I_{now}(p) \equiv p$ is inevitable now $\equiv p$ is inevitable, as the 'is' of A-theory is present-tense I retain the subscript to keep the A/B symmetry salient

only if the A-series truism implies

'E will occur at t' is true $\supset$ I_{now}(E will occur at t),

which it does not. Likewise, I_{t^-}(E occurs at t) follows from the truth at t^- of 'E occurs at t' and the B-series truism

I_{t^-}('E occurs at t' is true at $t^- \supset$ E occurs at t)[23]

only if the B-series truism implies

'E occurs at t' is true at $t^- \supset I_{t^-}$(E occurs at t),

which it does not. 'E at t exists at t^-', when coherent, is insufficient as well as unnecessary for fatalism.

Piercing this B-theory murk, simultaneity forks give "the present existence of the future" a clear meaning. And compatibilists grant that the late tip of a fork certainly appears inevitable at its early tip.

III

But here the compatibilist's concessions end. Despite agreeing that R at (0, 2) is the late tip of a fork whose early tip is P at (0, 0), he declines to conclude that during the interval $[t = 0, t = 2)$, I cannot prevent my run, that at $t = 2$, I will (do) not run freely, or that in retrospect at $t = 3$, I could not have done otherwise. The existence of PVR, he calmly repeats, implies nothing about the cause of R at $t = 2$ (that is, at $t' = 0$),[24] so the existence of PVR permits preference P to cause R. On the plausible assumption that P does cause R, he continues, I did R (at least weakly[25]) freely. On the equally plausible assumption that I would not run if I did not want to, I could have done otherwise.

The compatibilist cites two features of the Lorentz transformation to explain how an event at the early tip of a simultaneity fork can cause an event at its late tip, he holds in reserve a third feature that bears on the counterfactual [~I want to run → ~I run]. The first is light cone invariance: an event E^- in the rear (forward) light cone of event E^+ in one frame is in its rear (forward) light cone in all frames. Suppose that, in some arbitrary (x, t) frame, E^- at (x^-, t^-) is in the rear light cone of E^+ at (x^+, t^+), so $|x^+ - x^-| < |t^+ - t^-|$ and $\sqrt{\Delta x^2 - \Delta t^2}$ is imaginary. As separation is invariant, $\sqrt{\Delta x'^2 - \Delta t'^2}$ is also

[23] To this truism Nozick opposes quantum theory—see *Invariances* (Cambridge: Harvard, 2002), pp. 29ff. The truism is consistent with his later argument from SR to the relativity of truth.

[24] That is, at the worldpoint at (0, 2) in F and (2.65, 0) in F', the frame with origin (-3.33, 0) in F.

[25] See footnote 11(4).

imaginary in any other frame (x', t'), that is, E^- is absolutely in the light cone of E^+ Reverting to Figure 2, as P is in R's rear cone in F, P is in R's rear cone in F' and all other frames To be sure, the spatial and temporal distances between P and R contravary with frame The Lorentz map carries (0, 0) to (4 15, –2 5) and (0, 2) to (2 65, 0), so in F I start to run where I start to want to run, while in F' I start to run 1 5 lighthours from where (and 2 5 hours from when) I start to want to run Nonetheless, want precedes run in all frames, so the sequence <P, R> meets the customary condition that a cause must (absolutely) precede its effect

Forking is consistent with light cone invariance because there is no frame in which a fork's vertex is simultaneous with both its tips Although P and R both occur at the time of V, P occurs at the time of V in F, while R occurs at the time of V in F' That there is no frame in which V occurs at the time of P and of R is assured by the second pertinent feature of the Lorentz map no frame assigns the same time to events whose proper time difference is nonzero (This also justifies the definition of [absolutely] early and late tips) P precedes R by two hours in their proper office frame F While P is simultaneous with V in F, P precedes V by 2 5 hours in F' SR allows causes to precede their effects invariantly by preserving proper time order

Forks allow behavior to be caused by (invariantly prior) preferences, according to compatibilism the everyday criterion for the use of "free " It appears that relativistic fatalism must recognize the possibility of freedom in at least the compatibilist sense

IV

A relativistic fatalist will claim to be able to absorb causation by preference[26] and all the compatibilist's pet points about ordinary usage "So be it," the relativistic fatalist will say, "'Freedom' as ordinarily used does not require a causally indeterminate future, or a future the agent believes is causally indeterminate, only a future caused in part by the agent's wants The core use of 'cause' may well be to denote energy flow, and no doubt energy flows unidirectionally from wants to actions in orientable space-times[27] The fact remains that

[26] The efficacy he grants desire does not amount to weak freedom An act is weakly free when the desire to do it preempts an exogenous back-up, relativistic fatalism does not assert the omnipresence of back-ups Also, the effect of a back-up would typically differ from that of the pre-emptive desire Had *S* wanted to leave his Locked room he would have stayed angrily, the "same" as staying placidly only because act individuation is coarse Relativistic fatalism claims that all things must go *exactly* as they do

[27] Superluminal velocities create anomalies, as do time loops However, compatibilism is untroubled by consistent loops, such as a run that causes an earlier desire that causes the run Compatibilists can declare such a globally self-caused run free because

your so-called 'future' is a done deal As we speak, energy is flowing into 'future' time-slices of your quadriceps No matter how the situation is described, you cannot at noon your time refrain from running at 2 your time, and looking back from 3 your time you could not have done otherwise, because you cannot refrain from what you are already doing You compatibilists set no store by the categorical power to do otherwise, period, but the counterfactual conditionals you do think necessary for freedom are uniformly false Lockwood fights shy of these counterfactuals, but Rietdijk boldly denied them

> I could not possibly have influenced event [R] in an arbitrary way (e g , have prevented [R]) at any moment when [R] was still in the future, or was present, for me at [O], supposed that I did desire to do so (*op cit* , p 342) "

Thus speaks the relativistic fatalist

We should pause for a moment to consider how extraordinary a construction this is to put on even a revolutionary scientific theory Can SR really require we believe that I would run at $t = 2$ even if all I wanted at $t = 0$ was to grade at $t = 2$, or that if at $t = 1$ I burn my running shoes a simultaneity fork will provide another pair? Thankfully, SR turns out not to imply this A somewhat less far-fetched construction, perhaps Rietdijk's despite his ipsissima verba, is that while there are events that would keep me from running if they occurred, my already running prevents them My burning my shoes would stop my running, but because of the fork my shoes cannot be burned SR does not require we believe this, either

---

its local cause is a desire Where events in their own light cones might come from in the first place is a natural but quite general question whose answer presumably covers self-caused acts The answer will not cover paradoxical loops, such as a run causing an earlier injury that prevents the run, but no one need worry about them, because they are logically impossible John Earman—*A Primer of Determinism* (Boston Reidel, 1986), pp 175ff—claims to know of no good argument that self-canceling events are logically impossible, the following argument comes to mind If there could be a space-time path along which an event cancels itself, there could be an event that occurs iff it does not, that is, an event that does and does not occur Since $\sim\Diamond(p \wedge \sim p)$, no such path is possible A man can perhaps aspire to run fast enough to have hobbled himself earlier and believe he has succeeded, just as he can aspire to be and believe he is Russell's barber or the set of all sets, but success is not an option Gödelian universes exclude self-canceling events in the same way a shaver of all non-self-shavers is kept from becoming Russell's barber when he takes a razor to his own chin The hard problem, so to speak, is why we wonder what mechanism blocks self-canceling time travel but not what mechanism prevents Russell barbering Noson Yanofsky puts time travel paradoxes with the standard ones in a general setting in "A Universal Approach to Self-Referential Paradoxes, Incompleteness and Fixed Points," *Bulletin of Symbolic Logic*, IX, 3 (2003) 376–77

The idea that simultaneity forks in some manner constrain events is, I submit, the pivotal error of relativistic fatalism It is an error because of the third compatibilism-friendly feature of the Lorentz transformation the map $(x, y, v) \rightarrow ((x-vt)/\sqrt{1-v^2}, (t-vx)/\sqrt{1-v^2})$ from F-coordinates to F'-coordinates is kinematics, not dynamics The map relabels worldpoints and the events at them without tracking any physical process It no more represents a constraint on events in space-time than the map $(\omega, r) \rightarrow rd\omega/dt$ from angular velocity to tangential acceleration represents a constraint on rotary motion A more thoroughgoing analogy for the relation of $\Delta t = T(\mathrm{R}, F) - T(\mathrm{P}, F) = 2$ to $\Delta t' = T(R, F') - T(P, F') = 0$ is provided by the projections of the vector L from New York to Boston under various coordinatizations of the Earth's surface When the Equator is the x axis and the Prime Meridian is the y axis, the length $|L_x|$ of the projection L_x is 90 miles When x' is an arbitrary great circle and y' is a great circle parallel to L, $|L_{x'}|$ is 0 Shifting from the (x, y) to the (x', y') system "transforms away" the difference in longitude between New York and Boston, but not by revealing any feature of L hidden in the (x, y) system, least of all that Boston is really as westerly as New York The shift from (x, t) to (x', t') transforms away the time difference between P and R similarly

Changes in axial components under changes of axes do not prevent, cause, necessitate, force, or explain events Just as there is nothing that cannot (or must) exist in New York, Boston, or anywhere else because there is a pair of axes that assign New York and Boston the same longitude, there is nothing that cannot (or must) happen at any worldpoint, including those of P and R, because there is a pair of pairs of axes one of which assigns V and P the same time, and the other of which assigns V and R the same time That space and time compose a Lorentz-invariant vector is undeniably an empirical fact—one we find very surprising—and the contravariation of these components undeniably plays a role in explanation However, this explanatory role must be construed properly Take the often used example of pion decay The proper lifetime $\Delta\tau$ of a pion, explained by the properties of its constituent quarks, is sufficiently great that the vector from the creation of a pion in the Earth's atmosphere to its decay exceeds the vector from its creation to arrival at sea level (in both Earth and pion frames) Pions do not reach sea level because "motion dilates time," an interpretation of SR that physics texts regularly warn against Where t is proper Earth time and Δt the time to decay in the Earth frame, the $\tau \rightarrow t$ transformation explains why $\Delta\tau < \Delta t$, that is, why the time component of the creation-decay vector is greater in the Earth than in the pion

frame [28] Another example a charge passing a current experiences a force because of the difference, in the charge's rest frame, between the positive and negative charge densities of the carrier, not because "motion contracts space", what the Lorentz transformation explains is the geometry of the deflection in the observer frame This is not to deny that events are explained by relativistic dynamics (Compton scattering by the increase of mass with velocity, for instance), or that the dynamics is closely tied to the kinematics The point, rather, is that only the dynamics explains events, and follows from the kinematics only given an independent—and to us surprising—dynamical assumption, usually expressed as the generalized conservation of momentum or the invariance of the energy-momentum 4-vector under the Lorentz map The purely kinematical character of the map itself may be put as follows the occurrence of an event E at a worldpoint p,[29] whose coordinates in a frame F are by definition $(X(p, F), T(p, F))$, is never explained by the existence of a frame F' such that $X(p, F) = X(p, F')(\sqrt{1-v^2})-vT(p, F)$ and $T(p, F) = T(p, F')\sqrt{1-v^2})-vX(p, F)$, or, equivalently, by the existence of an F'such that $X(\mathrm{E}, F) = X(\mathrm{E}, F')(\sqrt{1-v^2})-vT(\mathrm{E}, F)$ and $T(\mathrm{E}, F) = T(\mathrm{E}, F')\sqrt{1-v^2})-vX(\mathrm{E}, F)$ More loosely, changes in the coordinates of events E_1, , E_n under changes of axes never explain the occurrence or functional relations of the E_i [30] I would emphasize that relativistic fatalism nowhere appeals to relativistic dynamics (see footnote 1) [31]

By this point, one hopes, the relativistic fatalist will have granted that event P, the onset of my desire to run at a worldpoint with t-coordinate 0 and t'-coordinate −2 5, may cause event R, my running, at a worldpoint with t-coordinate 2 and t'-coordinate 0 Let us assume P does cause R, there is no need to specify the cause of event V, the presence of e_v ($v = dx/dt = 6$) at that worldpoint with x-coordinate 3 33 and $t = t' = 0$ The simultaneity of V with R in F' is not why R will

[28] The $\tau \rightarrow t$ transformation does not explain observations of Δt and $\Delta\tau$, events explained by the action of pions, mediated by instruments, on human sense organs

[29] 'At p' is redundant if an event's identity includes its worldpoint

[30] Quineans, who gauge the equivalence of descriptions by intertranslatability and find translation indeterminate, seek to blur the line between pairs of equivalent descriptions and pairs of rival theories Space-time frames are determinately intertranslatable

[31] Confusing kinematics with dynamics leads to needless perplexities in other areas For instance, it sustains the prevalent view of Arrow's Theorem as a threat to democracy The Theorem states that, given individual preference rankings R_1, , R_n that meet plausible rationality conditions, any "social" ranking $R = f(R_1, , R_n)$ that meets further plausible conditions must be one of the R_i This i is customarily called a "dictator," an epithet implying that i causes the ranking R to coincide with his own The Theorem supports no such interpretation Indeed, it permits the dictator to change as the R_j do The cash value of the Theorem is simply that f always coincides with some projection function on $<R_1, , R_n>$ My thanks here to George Gramos and Andrew Mischenko

occur in 2 hours (from the perspective of O) or (from the perspective of O′) why R is occurring now That the time coordinate of R in F' is 0 does not force or necessitate or explain the occurrence of R at $t = 2$ in F, R at $t = 2$ is necessitated, if one favors that idiom, by P at $t = 0$ (and at $t' = -2.5$) R's worldpoint, being (2.65, 0) in F', must by relativistic kinematics be (0, 2) in F, and vice-versa, the kinematics demands nothing more As far as fork PVR is concerned—given what we ordinarily believe and are presently assuming about the causal relations between desire for exercise, exercise, and footware—I can at $t = 0$, the time of V in F, keep V from being simultaneous with R R will (does) not occur at $t' = 0$, and there will be (is) no R for V to be simultaneous with in F', if at $t = 0$ I commit myself to grading all afternoon The example stipulates that I *will not* stop V from being simultaneous with R, so, by stipulation, *will not* commit myself to grade I *cannot* commit myself to grade only if my running is (stipulated to be) compulsive With or without the compulsion, the fork does not forbid my wanting to grade

SR admits a distinction between what is and is not up to me that parallels the distinction as drawn by common sense and compatibilism Take the perspective of V, simultaneous with R in F' because of the formation of preference P P does not form, nor do I run, because V is simultaneous with R in F' Fork PVR exists because of an event at (0, 0) in F, that is at (4.15, –2.5) in F', not vice-versa Contrary to Rietdijk's bold assurance that I could not have prevented R had I wanted to, V would not have been simultaneous with R in F' had I at $t = 0$ ($t' = -2.5$) wanted to grade rather than run [32] Whether I run at $t = 2$ is in this conditional sense up to me at $t = 0$, and in this same conditional sense I can prevent my running, despite the impalement of my running on one tine of a fork *whose existence depends on my wanting to run* V *is* simultaneous with R in F', and I *do not* prevent R—but not because $T(V, F') = T(R, F')$ By contrast, unless I have unusual powers, V would have occurred no matter what I wanted, so in this conditional sense, V is at no point up to me

Relativistic kinematics likewise allows change of mind, holding V and P fixed, I could have intervened after P to avert R This we know fork PVR exists, and [I burn my running shoes at $t = 1$ ($t' = -1.25$) $\supset$ ~PVR exists], so I do not burn my running shoes at $t = 1$ Suppose for convenience that intact shoes also suffice for running, hence that

[32] Denial that my grading is free unless my wanting to grade is free strikes compatibilists as a conflation of freedom to do with freedom to want, see footnote 11(3) Relativistic fatalists must anyway forego responses that imply satisfiable sufficient conditions for freedom

[~I burn my shoes at $t = 1 \equiv$ PVR exists] The question is, Which explains which, do I spare my shoes because of the fork, or is there a fork because I spare my shoes? (Relativistic fatalists cannot afford appeal to a common cause [33]) The existence of the fork amounts to the constancy of the time coordinate of V and the change in the time coordinate of R (and the time coordinates of P and Q) in the shift from the F axes to the F' axes As we have decided that changes of coordinates under change of axes do not explain events, the existence of the fork does not explain my sparing my shoes It follows that my sparing my shoes explains the fork,[34] and that the fork would not exist unless I spared them Even with V and P held fixed, then, there would be no R and no fork were I to burn my shoes Since the fork and my running depend on and are explained by my sparing my shoes at $t = 1$, in hindsight at $t = 3$ I could have refrained from running by burning my shoes at $t = 1$ I *will not* burn them—by stipulation, not because I am already using them I *cannot* burn them only if I am neurotically averse to damaging sports gear

Despite packing a different physical theory, relativistic fatalists follow the trail to confusion blazed by their classical predecessors Where '$I_{t,F}(p)$' is 'p is inevitable at t in F', relativistic fatalists seek to reach $I_{0,F}[T(R, F) = 2]$[35] from $T(R, F') = 0$ and the kinematical lemma

$$I_{0,F}[T(R, F') = 0 \supset T(R, F) = 2]$$

But $I_{0,F}(T(R, F) = 2)$ follows only if the lemma implies

$$T(R, F') = 0 \supset I_{0,F}[T(R, F) = 2],$$

which it does not All $T(R, F') = 0$ implies is $T(R, F) = 2$ $I_{0,F}[T(R, F) = 2]$ looks plausible because PVR seems to make R's simultaneity with V in F' a fact—a reality, something that obtains—at $t = 0$ in F On closer inspection, however, 'R is simultaneous with V at $t = 0$' is no better defined in SR than was 'E at t exists at $\bar{t}$' in B-theory, and for a kindred reason there is no frame for the 'is' of 'R is simultaneous with V at $t = 0$' R and V occur at the same time in F', but the simultaneity of R and V in F' is not itself an event, so has no time in any frame '$T(V, F') = T(R, F') = T(V, F) = T(P, F)$' is of course true at $t = 0$ in F, as it is at all times in all frames (see footnote 21), but

[33] A common cause for the sparing and the fork vitiates necessitation of the sparing by the fork In fact, a natural cause of both the sparing and the fork is my preference

[34] Pragmatically, nonoccurrences explain only when there is some presumption of occurrence Imagine that I think long and hard about setting fire to my shoes

[35] Making the goal $(\exists E)(E = R \wedge I_{0\,F}T(E, F) = 2)$ brings it no closer

trying to reach fatalism from the truth of 'T(V, F') = T(R, F') = T(P, F)' only leads deeper into the B-theory jungle

Indeed, relativistic fatalists would be the first to insist, correctly, that the rule

Nothing at t can affect what has happened by t

must be relativized to frames But the correctly relativized rule

No event E at T(E, F) in F can affect what has happened by T(E, F) in F

lets events at T(E, F) in F affect events E' at T(E', F') $\leq T$(E, F) in $F' \neq F$, hence lets events at T(E', F') in F' depend on events at T(E, F) $\geq T$(E', F') in F (As T(E, F) is a dimensionless number, 'T(–, F') in F' is well defined) In particular, the relativized rule lets an event at $t' = 0$ depend on events at $t = 0$ and $t = 1$, so does not deny that, at $t = 2$, I could at $t = 0$ have failed to be about to run at $t' = 0$, or affirm that I would have been about to run at $t' = 0$ (even) had I changed my mind at $t = 0$ and burned my shoes at $t = 1$

You can see why compatibilists have no dog in the fight about the present reality of the future They can agree that simultaneity is transitive across frames and that an event is real at the times of all the events with which it is simultaneous, and add that all this means is that events can be prevented when and after they are real Preventing is the flip side of explaining, a nonevent E, such as persistence of my desire to grade, is prevented by what explains its nonoccurrence, an event E, such as a run, is preventable by any (non)event that, were it to occur, would explain the nonoccurrence of E A natural necessary condition for E's explaining (or potentially explaining) E's nonoccurrence is that, given E', E cannot happen Now, the simultaneity of V with R in F' and the existence of fork PVR are explained by three factors the onset at $x = 0$, $t = 0$ ($x' = 4\ 15$, $t' = -2\ 5$) of a desire to run in $\Delta t = 2$ ($\Delta t' = 2\ 5$) hours, the survival of my shoes after $t = 1$, and whatever brought e_v to $x = -3\ 33$, $t = 0$ These explanations imply that the desire could have existed without the fork, indeed would have existed without the fork had I burned my shoes Since the fork thus fails a necessary condition for explaining why I do not change my mind, it does not prevent a run-preventing change of mind at any time in any frame R can be real at $t = 0$ and, consistently with the relativized rule, preventable (by a change of mind) at $t \geq 0$ As for the eldritch-sounding "Events can be prevented after they are real," compatibilists blame it entirely on relativistic fatalists, the ones who insist that a fork with early tip E at t and late tip E^+ at t^+ entails the reality of E^+ at t That insistence, plus the concession that E^+ may nonetheless depend on E, implies that E^+ may be real

at t and preventable up to t^+ [36] Here is where the advantages of compatibilism over libertarianism may be most clear It is hard to see how a libertarian can posit a plurality of possible futures, or deploy his familiar metaphors of branching paths and tree trunks gradually denuded of branches, and yet grant that in *any* sense one future is presently real

Keats's *Ode on a Grecian Urn* has inspired comparison of spacetime to a continuum of urns (with or without the natural Newtonian metric) on which figures pose in eternal stasis Compatibilists, unfazed, consider the figures free insofar as their postures on one urn result from their desires on another The lover reaching for the maid on urn u embraces her freely on urn u^+ if he does so because on u he yearns to With so many urns the maid will be faded on some and absent from others, hardly what Keats had in mind, but mortality is a small price for a kiss, and freedom

V

There might seem little left to say about the bearing of SR on free will The compatibilist has sought to square SR with my unexercised power to grade by arguing that I would (have) grade(d) if I (had) wanted to, fork PVR notwithstanding In so arguing he has assumed his signature, and highly controversial, analysis of

(1) *S* can do B

as the conditional

(2) *S* wants to do B → *S* does B

The issue now becomes the adequacy of this analysis, and SR, as an empirical theory, appears ill equipped to call this analysis into question Indeed, as the relativistic fatalist bases his denial of the power to do otherwise on the falsity of all statements of form (2), he at least has a certain investment in (1) implying (2)

The situation, however, is more complicated There are two basic objections to the conditional analysis of power the supposed categoricity of (1), and the supposed difficulties of interpreting (2) As we will see in section VI, SR helps meet the second objection, in this way coming to the aid of the conditional analysis, and, surprisingly and

[36] "The difficulty we created here arose because we effectively equated coexistence with simultaneity The latter is a concisely defined mathematical condition, while the former implicitly carries much wider connotations If, however, one takes the message of [SR] to be that there is no absolute distinction between past and future, a person can (but need not) arrive at the so-called 'block-universe' model in which existence is seen as deterministic Such a view would not seem to be essentially different from the Laplacian determinism of classical physics"—James T Cushing, *Philosophical Concepts in Physics* (New York Cambridge, 1998), p 250, see fn 17

ironically, free will The first objection cannot be met by appeal to SR, but is so basic that it must be addressed, if only briefly, less further discussion seem unmotivated

Many critics will deny that it matters how a merely conditional power to do otherwise fares under SR, for freedom, they say, demands the categorical power to do otherwise *period, given everything as it is* Imagine S about to do A at t (and going on to do A), with @ the actual past up to t plus the actual laws of nature For S to do A freely, these critics say, it is essential that, for some B $\neq$A,

(3) S can do B at t, given @,

whereas d-compatibilists deny

(4) $\Diamond$(@ $\wedge$ S does B)

These critics may therefore think they do not care whether the compatibilist power to do otherwise survives SR, since "saving" freedom by enfeebling it only hastens its death

They should care, since conditional power may also be categorical The two cannot be reconciled if (3) is equivalent to (4), but this equivalence is open to doubt By (3) the guardians of categoricity plainly intend to assert that the world is (of course) as it is *and yet* S can do B They must therefore intend (3) to assert, or at least imply, perhaps with help from additional but evident premises, both @ and (1) (4) does imply

(5) $\Diamond$(S does B),

which incompatibilists take to be (1), but there is no way to get @ from (4) Hence (4) is too weak to be (3) Note, though, that both @ and (1) do follow from

(6) @ $\wedge$ (1)

Moreover, when (1) is construed along incompatibilist lines as (5), (6) becomes

(7) @ $\wedge$ (5),

which implies (4) (6) is therefore a reasonable explication of (3) More speculatively,[37] where π is a past with our laws[38] perhaps differing from @ (in initial conditions),

(8) If S does A in π, S could [still] do B given π

[37] The compatibilist does not need analyses of (8) and (10) to secure the categoricity of actual powers But note that [(9) $\wedge$ (11) $\wedge$ π = @] implies (6)

[38] In anticipation of the backtracking in section VI, the laws of @ and π are assumed identical

might be explicated as

(9) $(\pi \wedge S \text{ does A}) \supset (\pi \wedge (1))$,

and

(10) If *S* were to do A in π, *S* would [still] be able to do B given π

as

(11) $(\pi \wedge S \text{ does A}) \rightarrow (\pi \wedge (1))$

But compatibilists can also accept (6) and thus (3), which is categoricity in primitive notation To begin with, all sides may accept (2) taken on its own, apart from whether it is equivalent to (1) Consequently, all sides may accept

(12) @ $\wedge$ (2)

taken on its own, apart from whether (12) is the coveted power to do otherwise And as (12) does not imply (4), (12) does not flout determinism At the same time, compatibilists construe (1) as (2), and therefore (6) as (12), making (12)—something all sides may agree to, and consistent with determinism—the compatibilist interpretation of (3) One might add that compatibilists construe (9) as

(13) $(\pi \wedge S \text{ does A}) \supset [\pi \wedge (S \text{ wants to do B} \rightarrow S \text{ does B})]$

and (11) as

(14) $(\pi \wedge S \text{ does A}) \rightarrow [\pi \wedge (S \text{ wants to do B} \rightarrow S \text{ does B})]$,

since all sides may accept (13) and (14), and neither flouts determinism, (13) is a suitable compatibilist interpretation of (8), and (14) of (10) The compatibilist has delivered categoricity in power [39] Nor has he begged any questions on the way the claim was that his analysis can not deliver categoricity, entitling him to assume his analysis, as he did in equating (1) with (2), to show that it can [40]

[39] Austin's misreading in "Ifs and Cans" of the compatibilist analysans as '*S* could do B if he wanted', apart from erecting a straw man by making the analysis blatantly circular, abets the objection that *S*'s power in some unactual world is irrelevant to his power in this one—see, for example, Fischer, "Frankfurt-Style Compatibilism," p 24, fn 28 The compatibilist analysans is '*S would* do B if he wanted', which judges *S*'s power in this world by what he *does* in another one

[40] Ted Warfield's disproof of compatibilism—"A New Argument for Incompatibilism," *Philosophical Perspectives*, XIV (2000) 167–80—falters over confusion of (3) with (4) Suppose @ is deterministic, *S* does *A* in @, but, as per compatibilism, *S* could have done B anyway Given determinism, $\Box(@ \supset \sim B)$ So had *S* done B, Warfield argues, ~B would have happened, because $\Box(@ \supset \sim B)$ entails $\Box((@ \wedge B) \supset \sim B)$ Yet had *S* done B, B would have happened, because $\Box((@ \wedge B) \supset B)$ is logically true But surely,

Two further considerations bolster this argument First, it has a firm intuitive basis in the logical equivalence of ($p \wedge$True) to p Since everything always is as it is, (p, given that everything is as it is) is always equivalent to p,[41] even when p is a conditional whose antecedent is false In particular, 'S would do B if he wanted to' is equivalent to 'S would do B if he wanted to, given that everything is as it is (including his not wanting to)' So if 'S can do B' means 'S would do B if he wanted to', 'S would do B if he wanted to' trivially entails 'S can do B, given that everything is as it is'

Second, compatibilists can deliver categoricity without equating (3) with (6) Noncircular arguments like the one presented yield 'S can do B, given @' and 'If S does A in π he could do B, given π' when (3) is explicated as [@ $\rightarrow$ S can B], or Prob(S can B/@) > 0 Indeed, all compatibilists really need is the distinction between (6) and (4), for their key substantive claim is that, whatever paraphrase of (3) anyone may say he accepts, including as (4), he will ascribe power, assess responsibility, and sanction punishment as if he meant (6) Compatibilist preservation of categoricity is robust

At the same time, the argument does not pretend to close the basic rift dividing incompatibilists from compatibilists Incompatibilists still insist, and d-compatibilists deny, that [S does A freely $\supset \Diamond$(@ $\wedge$ S does B)] The argument purports only to display *a* sense of "could do otherwise, period," or rather, per the second point, several senses, available to compatibilists Perhaps one of these senses has begun to sound natural to you, and the incompatibilist sense strained, if so, advantage compatibilism But many philosophers will remain uneasy Even if 'wants to do $\rightarrow$ does' is in some sense(s) "categorical," they will continue to doubt its adequacy as an analysis of 'can' especially insofar as "can do otherwise" is necessary for "does freely " With that disquiet we resume the discussion of SR broken off earlier in this section

VI

There is a popular argument, couched in "possible worlds," that articulates many of the qualms felt about the conditional analysis of

---

Warfield protests, $\Box(p \supset q)$ should exclude $\Box(p \supset \sim q)$ (Here p = (@ $\wedge$ B) and q = ~B) In fact, $\Box(p \supset q)$ excludes $\Box(p \supset \sim q)$ only if $\Diamond p$, a condition not met in the present case (@$\wedge$ B) is impossible by hypothesis Had S done *B*, given determinism, the laws and the past would have been such as not to entail ~B In saying that in @ S is free to do B, the d-compatibilist makes no commitments about what would happen if S did B in @

[41] '{w @ is true at w}' is rigid, '{w the past and laws of w are those of actuality}' is accidental To speak two-dimensionally, 'p' would entail 'p, given everything is as it is' whatever proposition 'Everything is as it is' expressed, that is, whatever '{w the past and laws of w are those of actuality}' denoted

power This argument uses a certain tacitly fixed parameter δ about possible pasts, in a way seen to be improper when, by reference to SR, δ is allowed to vary Ironically, SR thereby comes to the rescue of compatibilism and thus free will

The argument starts with *S*, as usual, poised at *t* to do A Compatibilists say he can nonetheless do B iff he does B at *t* in the nearest[42] possible world at which he wants to With their d-hats on, they add that *S* would do B only if he wanted to (or things differed from @ in some other way), that is, that *S* wants to do B at the nearest world where he does B, and can do B if that is the nearest world at which he wants to However, specifying the nearest world at which *S* does B is said to raise insuperable problems for compatibilists Should they say it is the nearest world that matches ours before *t*, obeys our laws after *t*, and at which *S* anomalously wants to do B at *t*, they imply, absurdly, that *S*'s doing B involves a "miracle "[43] Should they say it is the nearest world with our laws[44] whose past differs enough from ours for *S* to want to do B at *t*, they imply, absurdly, that *S*'s doing B triggers a reverse butterfly effect Tracking back from *S*'s wanting B to, for example, a different upbringing, requires that his parents' upbringings also differ, and their parents' until "the entire past [must] be substantially altered "[45] And since d-compatibilists hold that our laws plus our past implies ~(*S* does B), they must specify the world in question in one of these two ways

Compatibilists are afforded some wiggle room by the tu quoque character of this dilemma Libertarian incompatibilists for their part think *S* can *just* do B at *t*, that is, that *S* does B in some legal world with our past, so they do not care where—the laws or the past—a further factitious difference is discerned Under pressure they might

[42] I ignore the difficulties that arise if there is a tie for nearest, or actuality is a limit world

[43] A tendentious misappropriation from David Lewis Citing Lewis's rule that agreement in fact may outweigh difference in laws in measuring nearness, Jonathan Vogel deems the "apparently true" conditional 'I turn dolphins into men → I work a miracle' true on Lewis semantics—"Tracking, Closure and Inductive Knowledge," in Steven Luper-Foy, ed , *The Possibility of Knowledge* (Totowa, NJ Rowman and Littlefield, 1987), pp 197–215, on p 211 Miracles for Lewis cannot occur by definition A miracle violates the laws of the world at which it occurs, and laws are, by definition, exceptionless The laws of a world at which I change dolphins into men let me do so, so I work no miracle See Lewis, "Counterfactual Dependence and Time's Arrow," in *Philosophical Papers II* (New York Cambridge, 1986), pp 32–52, on pp 44–45

[44] World *W* with past π and laws *L* is nearer to a world *W'* with past $\pi' \neq \pi$ and laws *L* than to a world *W"* with past π' and laws $L' \neq L$ So if *W* is actuality and π' includes *S*'s wanting B, our laws are the laws of the nearest world where *S* wants B

[45] Tomas Kapitan, "A Master Argument for Incompatibilism?" in Kane, ed , pp 127–57, on p 155 Kapitan accepts a version of compatibilism, so his remark is not wholly hostile, also see Kadri Vihvelin, "Libertarian Compatibilism," *Philosophical Perspectives*, XIV (2000) 139–60, on pp 154–55

grant that *S* would not do B unless he chose to, but promptly add that *S* can *just* choose to [46] Hard determinists do not think *S* can do B, but agree that *if S* could do B, he could *just* do it Impossibilists, too, agree that *S* can do A freely only if he can *just* do B, but also think B's occurrence in a world otherwise just like ours would make A's occurrence in our world random, hence unfree [47] One way or another, for incompatibilists the meaning of $\rightarrow$ is a nonissue [48] What is more, the extension of the dilemma I am about to present works equally well against miracles and backtracking This indifference to the semantics of $\rightarrow$ leaves d-compatibilists free to parse [*S* does $\rightarrow$] as they please Mostly for definiteness I will assume that they backtrack [49] They already think *S* wanted to do B at the nearest legal world where he does B, so I am assuming that they think this is the nearest world, period, where he does B In effect my d-compatibilist accepts [*S* does B $\rightarrow$ *S* wanted to B], [*S* wanted to B $\rightarrow$ *S* had upbringing U], [(*S* wanted to B $\wedge$ had upbringing U) $\rightarrow$], I add that, Janus-like, counterfactual antecedents may face two ways, letting d-compatibilists also accept foretrackers like [*S* wants to B $\rightarrow$ *S* does B] and [*S* does B $\rightarrow$ *S* feels regret] [50]

[46] Many libertarians take the more nuanced, and considerably stranger, view that an agent is responsible for a causally determined act so long as its determinants include some earlier undetermined act On this view, adults are responsible for what they do in virtue of choices made as children I had always assumed I misunderstood libertarians on this point until I found Robert Kane holding seventeen-year-old Dylan Klebold responsible for his part in the Columbine shooting because of Klebold's "own previous choices"—*A Contemporary Introduction to Free Will* (New York Oxford, 2005), p 122

[47] Impossibilists agree that a free act can be causally determined so long as the agent freely creates some of its determinants, but see the regress generated by this requirement as vicious Aristotle begins his discussion of the voluntary in *Nicomachean Ethics* III as a compatibilist ("the voluntary is that whose cause is within"), veers in Kane's direction, seems to sense the looming regress, and abruptly moves on

[48] See Timothy O'Connor, *Persons and Causes* (New York Oxford, 2000), p 16

[49] Backtrackers are already a fact of life inside, outside, and on the edge of free will territory Frankfurt cases involve them, see footnote 11(3) They show up in game theory Axelrod's theorem on the robustness of tit-for-tat uses the lemma (Tit-for-tat defects at $n \rightarrow$ Opponent defected at n–1) A backtracker resolves the lottery paradox "I don't know I'll lose the lottery because I wouldn't see winning coming" (Oddly, Nozick's (back)tracking account of inductive knowledge in *Philosophical Explanations* ((Cambridge Harvard, 1981), pp 222–27) does not mention lotteries) Scientists willingly use laws to backtrack, as witness Richard Feynman "According to our equations [flux = rate of change of charge], if you suddenly place a charge at some point, you had to carry it there from somewhere else"—*The Feynman Lectures on Physics* (Reading, MA Addison-Wesley, 1965), II, p 18-3 Laws are put to such a use below

[50] No doubt war would break out after Nixon pushed the button, but surely he would have wanted to push it Even Lewis puts the miraculous break between "actual past and counterfactual present" far enough into the counterfactual past to allow "an orderly transition" (*op cit*, p 40) Paul Horwich, Hartry Field, and Jonathan Bennett have remarked on Lewis's endorsement of mini-backtracking

The case against compatibilist backtracking, as put by Fischer, is that "'multiple pasts'—different pasts for each path I could take" is an "unintuitive" and "unlovely" idea [51] Timothy O'Connor finds it "incredible" that "I can perform an action, the performance of which requires the entire past history of the world to have been different" (*op cit*, pp 16–17) Other incredulous incompatibilists will be cited presently For now I observe that, so far as I can tell, all of them presume classical Newtonian space-time as the cosmological matrix for backtracking Suppose they took seriously the Minkowskian character of the matrix Might they not smell blood in the ether as the consequences of hypothetical acts are sent ramifying back through interconnected space and time to create pasts even more divergent than classically? Might they then not argue that it is *more* incredible that S can perform an action, whose performance requires the past to have been more different than once imagined? Imagine S, about to do A, at rest before t in a room he would have had to cross to want to do B (Only the view from the window could have aroused this desire) On classical assumptions, S's mass in all frames at the nearest world in which he crosses the room is his actual rest mass Under SR, S's mass in a frame F at the nearest world in which he crosses the room differs from his actual mass in F as a nonlinear function of the difference between his actual and hypothetical velocity in F Laboring under this extra relativistic burden, the argument concludes, a conditional analysis of power looks uglier than ever [52]

I trust you need no convincing that this argument is flawed [53] For one thing, it is a special pleader Under SR, the antecedents of counterfactuals about inanimate things posit departures from reality as great as those posited by the antecedents of counterfactuals about agents, yet SR has not ended all counterfactual reasoning The mass in the room frame of a moving S is not only calculable from SR, one calculates it just as one calculates it classically One assumes S in motion, chooses some physical laws, and adjusts the factors the assumption and the laws say need adjusting (solving (or tabling) cotenability problems along the way) Whether the laws chosen are classical or their relativistic corrections makes no logical or metaphysical difference [54]

[51] Fischer, Review of Dennett, *Freedom Evolves*, this JOURNAL, C, 12 (December 2003) 632–37, see pp 635–36

[52] If S's wanting B entails the miraculous relocation of a particle in S's brain, part of SR's extra-miraculous miracle is a change in the propagation of nearby light

[53] Not every philosopher to whom I have broached it has rejected it out of hand

[54] This general procedure resembles structural modeling, whereby for each situation the relevant variables and their functional relations or fixed values are stipulated, see Christopher Hitchcock, "The Intransitivity of Causal Relations Revealed in Equations

The one difference made by choosing SR is epistemic SR requires adjusting more factors, which complicates computation Classical backtracking from arrival at the window to motion across the room uses S's actual rest mass for all frames, relativistic backtracking calls for recomputing S's mass for each frame as a function of relative velocity [55] That the (uniform) extra burden imposed by SR is computational,[56] and consequently falls on the investigator of counterfactuals rather than the agents (and things) they are about, points up the argument's second, more instructive, flaw By treating the burden as an imposition on agents, the argument conflates what would have to have happened for an agent to do otherwise with what the agent would have to do to do otherwise [57] For SR to deepen doubt about the involvement of different pasts in doing otherwise, the power to act in a way that requires a different past must be thought to dwindle as that past becomes harder to describe This inverse correlation is mystifying unless it is assumed, first (and perhaps reasonably), that the longer the list of differences between pasts the more they differ, but also that to do otherwise the agent would have to bring about (or bridge) the difference between the actual past and the requisite counterfactual one Conflation of requiring with bringing about is already nascent in Kapitan's talk of a "substantially altered" past [58] Who do counterfactuals ask to do the altering, and of what past? We who think about alternate pasts adjust only our opinions The agent who might do otherwise does not change our past, since he does not do otherwise in our world, nor does he change the past of the world of his doing otherwise, which is over before he acts The proper inference is not the impossibility of what counterfactuals ask, but that counterfactuals never ask their subjects to change the past or work miracles Well, hardly ever The antecedent of 'S changes the past $\rightarrow$ ' asks S to change the past, and that of 'S works miracles $\rightarrow$ ' asks S to work miracles Yet even these very singular counterfactuals remain subject to the require/cause distinction The

---

and Graphs," this JOURNAL, XCVIII, 6 (June 2001) 273–99, for criticism, see Ned Hall and Laurie A Paul, "Causation and Pre-emption," in Peter Clark and Katherine Hawley, eds , *Philosophy of Science Today* (New York Oxford, 2003), pp 100–30 While the versions of structural modeling known to me stipulate it away, backtracking is consistent with a modeling format

[55] A change in light propogation near the bullet is part of the relativistic miracle

[56] Randy Dipert suggested this way of putting it

[57] Lewis distinguished breaking laws from presupposing them broken, but not changing the past from presupposing it changed—"Are We Free to Break the Laws?" *Theoria*, XLVII (1981) 113–21 This omission probably reflects his resistance to backtracking

[58] Also in O'Connor's description of nonbacktracking compatibilists as "scofflaws "

way the past would have to differ for S to change the past may be broader than the way he changes it The miracle necessary for S to work miracles need not be a miracle he works

This second flaw gives the extra-burden argument a diagnostic use d-compatibilism implies that there is some nontrivial[59] net difference δ between the actual past and the past there would have been had S done B, δ is the parameter mentioned earlier While δ has a unique value, given a unique nearest world at which S does B, we have become aware of two values δ might be thought to take Backtracking via classical physics, actual and hypothetical pasts differ by $\delta = \delta_{\text{classical}}$, backtracking via SR, these pasts differ by the broader $\delta = \delta_{\text{relativistic}}$ I have explained why the extra-burden argument will persuade you only if you think [S does B $\rightarrow \delta$] entails [S does B $\rightarrow$ S brings δ about (or breaks a law)] The argument takes the shift from identifying δ as $\delta_{\text{classical}}$ to identifying δ as $\delta_{\text{relativistic}}$ as enlarging, by $\delta_{\text{relativistic}}-\delta_{\text{classical}}$, a gap S must create (or leap) to do B Since, barring threshold effects, more divergence between actual and hypothetical pasts gives S more to do only if divergence per se gives S something to do, the entailment is needed Conversely, the argument will persuade you if you accept the entailment For if doing B requires δ, where "require" implies "bring about," S's power to do B will depend on the scope of δ The more actual and hypothetical pasts diverge, the more S must *do* to do B, the harder doing B is, and the less likely S is to be able to do it Since $\delta_{\text{relativistic}}$ exceeds $\delta_{\text{classical}}$, acceptance of SR plus certainty that S *is* able to do B will deepen doubt that S's doing B involves any nontrivial δ Acceptance of the extra-burden argument, then, specifically and sensitively indicates blindness to the distinction between presupposing and bringing about

Such a test is needful because the importance of the distinction to d-compatibilists leads incredulous incompatibilists to belittle it, often with scorn bordering on outright denial Does this criticism cross the line into absurdity? According to d-compatibilists, I could not grade unless the past had been different enough for me to want to grade [60] Even if talk of "altering the past" is allowed, d-compatibilists do not require my grading to alter the past so long as whatever would have differentiated the past of a grading world from the actual past would not be my grading [61] In response, Wallace calls the distinction "fine

[59] A difference logically entailed by S's doing otherwise is trivial For instance, a world W at which S does B at t differs trivially from ours in that, at W, S at $t'<t$ will do B at t

[60] Or my grading were anomalous

[61] Beebee observes that 'S's doing B would take a miracle' lets S's doing B itself be the miracle Lewis's version of the cause/require distinction is open to this objection

and rather technical" and professes to find it "hard to believe that so fine a distinction could bear such heavy moral weight." Fischer also calls the distinction "fine," and "hard to explain and justify." Ekstrom objects that on the compatibilist analysis "what an agent now *can* do includes something that either itself abrogates or entails the abrogation (somehow or other) of the laws or the past," as if abrogating and entailing abrogation were on a par. O'Connor asks, rhetorically, whether it is "really any more incredible" that "I can perform an action, the performance of which would have brought about a series of events in the past [than that] I can perform an action, the performance of which requires the entire past history of the world to have been different." Ascending semantically, van Inwagen hovers over the require/cause line with a denial that anyone can "render false" any truth about the past.[62] Compatibilists will instantly point out the extreme selectiveness of this incredulity. Consider, instead of an agent, an inanimate object with a latent disposition, for instance a fragile but intact glass vase governed by deterministic laws. Something can now befall this vase, namely breakage, whose occurrence would require the entire past to have been different (backtracking from a hypothetical hammer blow), or render false a truth about the past ("No hammer struck this vase"), or entails abrogation of a law (by an anomalous blow). Yet (one trusts) no erstwhile incredulous incompatibilist will deny, at least for this reason, that "can break" implies "would break if struck."[63] Incompatibilists might seek to defend an implicit double standard for "can" by contrasting the active power of agents with the mere capacity of mere things to undergo change when

---

because it offers no analysis of 'could have done B'. The counterfactual analysis protects itself by putting B in the consequent of a conditional whose antecedent specifies the miraculous precursor. In the nearest world where *S* does B he does so from a prior want, not out of the blue; the miracle is the want, or the cause of the want. Strikingly, Vihvelin *defends* "libertarian compatibilism" on the grounds that an unperformed act can be the miracle at the nearest world of its performance.

[62] R. Jay Wallace, *Responsibility and the Moral Sentiments* (Cambridge: Harvard, 1998), p. 255; Fischer, "Freedom and Miracles," *Noûs*, xxii, 2 (1988): 235–52, p. 249; Laura Ekstrom, *Free Will* (Boulder, CO: Westview, 2000), p. 50; O'Connor, pp. 17–18; Peter van Inwagen, *An Essay on Free Will* (New York: Oxford, 1983), p. 68. That only some event or other must differ for me to grade still forces Lewis to say that my grading changes the past; Ekstrom argues (p. 47). The disjunction of events that might have occurred had I graded depends counterfactually on my grading, hence, on Lewis's counterfactual analysis of causation (and assuming events closed under disjunction), is an effect of my grading. This point reinforces the criticism that causation explains, hence cannot be analyzed as, counterfactual dependence.

[63] There are problems about finkish dispositions, akin to finkish disabilities; see footnote 11(7).

Columbia University Libraries

acted on Compatibilists acknowledge this contrast in some form, and agree that what is required for the exercise of active powers and what is required for the triggering of passive capacities do not in general coincide Exercising the power to read Kant requires a pro-attitude toward reading Kant, getting shattered requires no attitude pro or con However, compatibilists will insist, this contrast gains purchase in the present context only if "required for" the exercise of active power (but not the passive suffering of change) means "caused by " Accepting determinism for vases (I assume), incompatibilists must accept [The vase breaks $\rightarrow \delta^*$] for some nontrivial difference δ^* between the actual past and the past of the nearest broken-vase world But agreement that the vase could not break unless δ^*, plus denial that *S* could not do B unless δ, suggests the belief that doing B bears to δ some relation that breakage does not bear to δ^* What could this relation be but that of bringing about? Incompatibilists will protest that they are innocent of this non sequitur, that they recognize the require/cause distinction, and that all they dispute is its load-bearing strength To compatibilists this sounds like lip service Stalemate

This impasse might be broken by asking incredulous incompatibilists what they make of the extra burden argument, a question that presents them with a dilemma of their own Should they agree that SR strengthens the case against the conditional analysis of power, they confirm their conflation of "require" with "cause " If they demur, and admit that the existence of a power does not depend on how different the past must be for its exercise, they will find themselves at a loss to say what they cannot believe about powers whose exercise requires a different past An incompatibilist might seek to evade the dilemma by claiming to speak figuratively. In minimizing the require/cause distinction, he might explain, he does not seek to foist on compatibilism the view that free agents can literally change the past, merely to convey what he sees as the sheer Rube Goldberg convolutedness of the compatibilist analysis But this maneuver too would transform the dialectical battlefield d-compatibilists have staked out the position that [*S* does A freely $\supset$ (*S* does $\sim A \rightarrow \sim @$)] In an apparent attempt at a reductio, incompatibilists advert to the extravagances of backtracking (or miracle-mongering) If, to avoid submitting to the extra-burden test, incompatibilists disavow a literal reading of their own objection, they turn out not to have offered a counterargument at all They have *restated* in metaphorical terms their original tenet that [*S* does A freely $\supset (\exists B \neq A)\Diamond$(*S* does B $\wedge$ @)], with the addendum, if they are libertarians, that some *S*'s do some A's freely A double standard for the if-iness of abilities and liabilites would turn out to express the same unargued tenet by another trope, the existence

of en soi things and pour-soi initiators Marching in place, incompatibilists would not actually have counterattacked

Revival of interest in relativistic fatalism will prompt fresh looks at many of the issues surrounding free will Compatibilists expect the light thrown by SR on the conditional analysis of power to fall just as flatteringly on other features of their doctrine, revealing it to be even lovelier than realized

MICHAEL LEVIN

City College of New York and
Graduate Center-City University of New York

REASONS AS CAUSES IN BAYESIAN EPISTEMOLOGY*

In everyday matters, as well as in law, we allow that someone's reasons can be causes of her actions, and often are That correct reasoning accords with Bayesian principles is now so widely held in philosophy, psychology, computer science, and elsewhere that the contrary is beginning to seem obtuse, or at best quaint And that rational agents should learn about the world from energies striking sensory inputs—nerves in people—seems beyond question Even rats seem to recognize the difference between correlation and causation,[1] and accordingly make different inferences from passive observation than from interventions A few statisticians aside,[2] so do most of us To square these views with the demands of computability, increasing numbers of psychologists and others have embraced a particular formalization, causal Bayes nets, as an account of human reasoning about and to causal connections.[3] Such structures can be used by rational agents, including humans in so far as they are rational, to have degrees of belief in various conceptual contents, which they use to reason to expectations, which are realized or defeated by sensory inputs, which cause them to change their degrees of belief in other contents in accord with Bayes's Rule, or some generalization of it How is all of this supposed to be carried out?

I REPRESENTING CAUSAL STRUCTURES

The causal Bayes net framework adopted by a growing number of psychologists goes like this Our *representations* of causal relations are captured in a graphical causal model, or causal Bayes net We reason

*Thanks to Alison Gopnik for suggesting dynamic Bayes nets, and the James S McDonnell Foundation Causal Learning Collaborative for intellectual support

[1] Aaron P Blaisdell, Kosuke Sawa, Kenneth J Leising, and Michael R Waldmann, "Causal Reasoning in Rats," *Science*, CCCXI (2006) 1020–22

[2] For example, Karl Pearson, *The Grammar of Science* (Mineola, NY Dover, 2004) Originally published in 1892 and still in print, Pearson maintains that causation is correlation and that there is no fact to the matter of correlation or causation because what we experience is events in our brains

[3] A very limited sample, each with references to other relevant works, includes the following Clark Glymour, *The Mind's Arrows Bayes Nets and Graphical Causal Models in Psychology* (Cambridge MIT, 2002), Alison Gopnik, Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks, "A Theory of Causal Learning in Children Causal Maps and Bayes Nets," *Psychological Review*, CXI (2004) 3–32, Gopnik and Schulz, eds , *Causal Learning Psychology, Philosophy, and Computation* (New York Oxford, 2007), Steven A. Sloman, *Causal Models How People Think about the World and Its Alternatives* (New York Oxford, 2005)

0022-362X/07/0409/464–74 © 2007 The Journal of Philosophy, Inc

implicitly as though we were calculating explicitly (but often not quite accurately) with such a network in hand The network is a mathematical object describing relations among features of a system or situation that are potentially variable—for example, having at least *present* or *absent* as possible values Those features are vertices, or variables, in a network with directed edges from some vertices to others A set of conditional probabilities is associated with the network, specifying for each vertex, V, the probability of each of its values conditional on each specification of values of the vertices in the graph that are parents of V—that is, those that have edges directed into V The graph is almost always assumed to be acyclic there is no sequence of directed edges leading from a variable back to that same variable For example, a simple network relating a lamp to an electrical power source and a switch on a timer might be

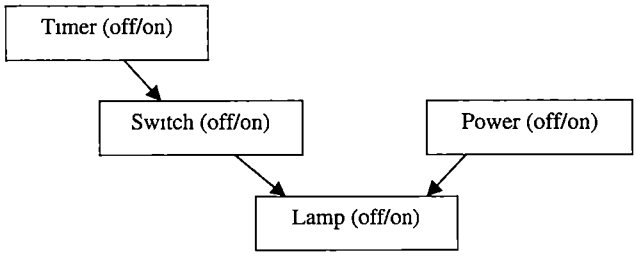


Figure 1 The Causal Bayes Net Ascribed to the World

Power and Switch have independent probability distributions The state of Lamp is determined uniquely by its two inputs Lamp is on if Power is on and Switch is on If the value of Power is ignored or unknown and varies from case to case, then the state of Lamp will appear to be an indeterministic function of Switch The causal content is captured by the supposition that a direct intervention that changes the state of a variable changes the state of variables downstream from it, but leaves the state of other variables unchanged [4] So, for example, an intervention that breaks the bulb of the Lamp leaves the state of the Lamp fixed at off, regardless of any variation in Power,

[4] A framework for specifying the consequences of interventions in a network is given in Peter Spirtes, Glymour, and Richard Scheines, *Causation, Prediction, and Search* (Berlin Springer, 1993) Algorithms for computing the effects of interventions that fix a definite value for one or more variables are given in Judea Pearl, *Causality Models, Reasoning, and Inference* (New York Cambridge, 2000), and discussed as a foundation for understanding causation by James Woodward, *Making Things Happen A Theory of Causal Explanation* (New York Oxford, 2003)

Switch, or Timer, and leaves the probabilities of Power, Switch, and Timer unaltered, as if the edges from Switch and Power to Lamp were broken by the intervention, though the arrow from Timer to Switch is left unaltered

The causal attribution carries with it expectations—in the human, not the probabilistic, sense—about the joint frequencies of events, and also expectations about the results of possible or hypothetical interventions on features of the system If an agent believes a light switch causes a lamp to go on, then the agent expects that turning the switch on and off will turn the lamplight on and off The agent can use the Bayes net to reason to a degree of belief in a particular event, or even to a specific predicted value for a variable, if other variables in the network have specified values if the light switch is believed to be chancy (but not overly so), the agent may derive a degree of belief that the lamp goes on when the switch is thrown, or may conclude that the lamp will go on In the other direction, the agent can use the Bayes net to reason backwards to alter her degrees of belief about the possible causes of some observed event, say that the lamp is on, in accord with Bayes's Rule If the agent's prior degree of belief that the switch is on is r, that the lamp light is on is p, and that the lamp light is on given that the switch is thrown is q, then after coming to believe that the lamp light is on, the agent's degree of belief that the switch is on changes to rq/p Further, the agent can use the Bayes net to reason hypothetically about the results of possible interventions that would fix one or more features from outside the system, and in particular, about the results of her or others' actions Suppose the timer is set to turn on the light switch automatically at a certain time The effect on the state of the lamp of an outside intervention that turns *off* the light switch at some later time is reasoned about hypothetically by supposing that the value of Switch is fixed in the Causal Bayes Net Ascribed to the World, but nothing else is altered In particular, the degrees of belief in the Timer state and that the lamp is on given that the light switch is off are left the same Switch, which formerly depended on its parent variable in the graph (Timer), now becomes independent (in the degree of belief measure) of its parent Graphically, the intervention breaks the directed edges from the parent variables—whatever they may be—to Switch The results are expected to match the causal consequences of the corresponding intervention in the world

This nice theoretical picture is substantiated by a variety of experiments that suggest that even young children make predictions and provide explanations that are patterned as a Bayes net requires, and change their confidence in outcomes roughly according to Bayes's Rule Some bits of the account are better established than

others, and of course people make errors and are computationally limited For example, no account is given of how people choose to attend to one phenomenon rather than another, and in psychological experiments that focus is almost always provided by the experimenter *Rector ex machina* Computer science aids the psychologists' account by providing a variety of algorithms for (comparatively) efficiently computing conditional probabilities in a Bayes net, and for computing probabilities given an intervention, that is, regardless of whether people make inferences just as the computer algorithms do, the inferences are at least feasible And, finally, recent work has shown that neural firing frequencies in a recurrent neural network—one with feedback loops—can implement an algorithm that computes some of the conditional probabilities defined in a Bayes net Moreover, the neural model corresponds well with firing frequencies observed in the visual cortex [5]

But can reasons like these, observations of features of the world that are causes or effects of other features, be causes? We do not mean to suggest somehow that the degrees of belief, or changes in them, are epiphenomenal and therefore not causal simply because the computations of conditional probabilities are carried out by neural processes, we are content with local identifications of changes in degrees of belief with instances of neural processes Our concern is rather with how and whether the reasoning that psychologists suppose agents do with the Causal Bayes Net Ascribed to the World can itself be consistently represented using a causal Bayes net, as should be possible if those reasons are causes

II CONNECTING CAUSAL BELIEFS AND INFERENCE

Let us assume (for the moment) that the connections and mechanisms needed for computing probabilities according to the Timer → Switch → Lamp ← Power network are somehow implemented in a reasoning agent Suppose now the agent wishes the lamp to light at 6 00 P M Her reasoning to a timer setting presumably goes something like this "If I set the timer for 6, then the switch will go on at 6 If the power is on at 6, then the lamp will certainly light at 6 The timer setting is independent of whether the power is on at 6 It is very probable that the power will be on at 6 Therefore, if I set the timer for 6, then the light will very probably go on at 6 "

[5] Rajesh R.N Rao, "Bayesian Inference and Attentional Modulation in the Visual Cortex," *Cognitive Neuroscience and Neuropsychology*, XVI (2005) 1843–48 See also Kenji Doya, Shin Ishii, Alexandre Pouget, and Rao, eds , *Bayesian Brain Probabilistic Approaches to Neural Coding* (Cambridge MIT, 2007)

So she sets the timer to go on at 6 00, and expects the lamp to go on at 6 00 Her reasons include both a desire and a sequence of degrees of belief about consequences of an action The reasons are causes, not only of her action, but also of the change in her degrees of belief that the switch will go on at 6 00 and that the lamp will go on at 6 00 As causes, her degree of belief reasons mirror the structure of the causal Bayes net structure she ascribes to the Timer/Switch/Lamp/ Power system, but the variables are now her own degrees of belief in various conceptual contents The goal that the light go on at 6 00, whether hypothetical or desired, somehow determines the relevant variables for the Causal Bayes Net Ascribed to the World (since there must be a great many such causal networks available to the agent), and the course of reasoning to the conditional forecast

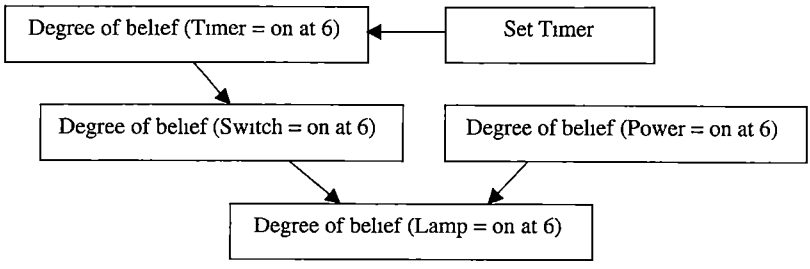

Degree of belief (Timer = on at 6)
Set Timer
Degree of belief (Switch = on at 6)
Degree of belief (Power = on at 6)
Degree of belief (Lamp = on at 6)

Figure 2 The Causal Bayes Net of Reasoning to a Forecast

The variables now range over values of degrees of belief—whether these are all real values between 0 and 1, or some finite range, as in high, medium, low, makes no difference here

The relations between degrees of belief in the Causal Bayes Net of Reasoning to a Forecast are chancy, as they describe some causal process in the brain, which may be subject to various chance fluctuations Accordingly, there are conditional probabilities associated with the directed graph, and in agreement with the psychological hypothesis that causes are represented as a graphical model, we will assume these conditional probabilities together determine a joint probability distribution [6]

Suppose now that the lamp does not light at 6 00, and the agent sees that it does not light According to the psychological story, she should then reason using the Causal Bayes Net Ascribed to the World by conditioning on Lamp = off to compute a new probability that the power

[6] We assume the chances satisfy the Markov property for the graph the joint probability distribution is the product of the conditional distribution of each variable in the graph given values of all its parent variables

is off, and a new probability that the Switch and/or Timer are off These new computations will constitute reasoning from the observations to new degrees of belief corresponding to new probability ascriptions, the observation about the lamp is a reason to change one's belief about other features of the world That picture seems eminently reasonable, any serious epistemological theory holds that we use observations to change our degrees or strengths of belief But that picture does not work with the Causal Bayes Net of Reasoning to a Forecast just described

The Causal Bayes Net of Reasoning to a Forecast specifies, prior to the agent perceiving at 6 00 that the Lamp is off, the causes of the agent's degree of belief that the Lamp is (or will be) on at 6 00 Those causes are her prior degrees of belief that the Switch is on at 6 00 and her prior degree of belief that the Power is on at 6 00, and more remotely, her prior degree of belief that the Timer is on at 6 00 In other words, perception of the light state is not a cause of degree of belief in the light state *from the point of view of this system.* Thus, the perception that the Lamp is off is an intervention on her Degree of belief that the Lamp is on, and so the perception that the Lamp is off at 6 00 cannot alter any of her degrees of belief in the other propositions, exactly because it is an intervention on Degree of belief(Lamp = on) in the Causal Bayes Net of Reasoning to a Forecast More formally, all of the edges into Degree of belief(Lamp = on) are broken, and so each pathway from other variables to Degree of belief(Lamp = on) is destroyed by the intervention Therefore, by the Markov property, all of the other Degree of belief variables are independent of Degree of belief(Lamp = on) after the perception But by Bayes's Rule, a proposition that is independent of a piece of evidence is not changed by acquiring the evidence, and the values of other nodes in the Causal Bayes Net of Reasoning to a Forecast are therefore unaltered by the perception of the lamp state

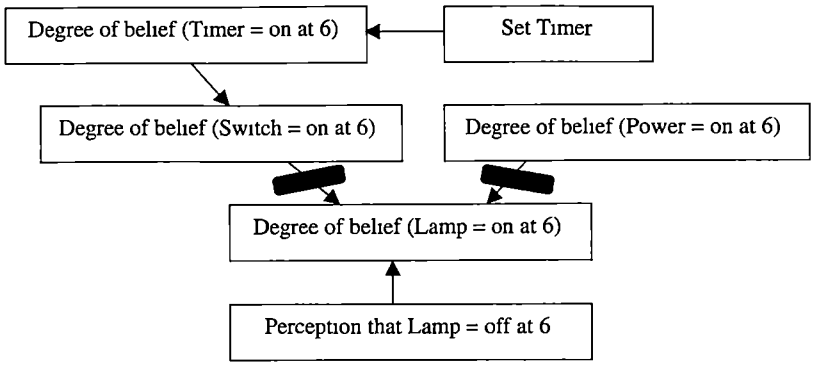

Degree of belief (Timer = on at 6)
Set Timer
Degree of belief (Switch = on at 6)
Degree of belief (Power = on at 6)
Degree of belief (Lamp = on at 6)
Perception that Lamp = off at 6

Figure 3 The Causal Bayes Net of Reasoning to a Forecast after Perception

The psychological story has a problem the view that degrees of belief, or changes in them, are causes seems incompatible with Bayesian learning from perception Perception of the state of an effect should lead (by Bayesian updating) to changes in beliefs about the causes, but perception is an exogenous intervention in the standard reasoning network, and so breaks the connections between the effect and its causes

Qualitatively, the agent's reasoning upon perceiving that the lamp is not lit at 6 00 goes something like this "The lamp is not on, therefore the probability that the power is on is decreased and the probability that the switch is on is decreased, because the probability that the switch is on has decreased, the probability that the timer is on is also decreased" The unmentioned sensation is the initial cause of this sequence of reasons If reasons are causes in this particular type of reasoning from perception to new degrees of belief in other contents, then the internal causal chain starting with the new degree of belief that the Lamp is on goes like this

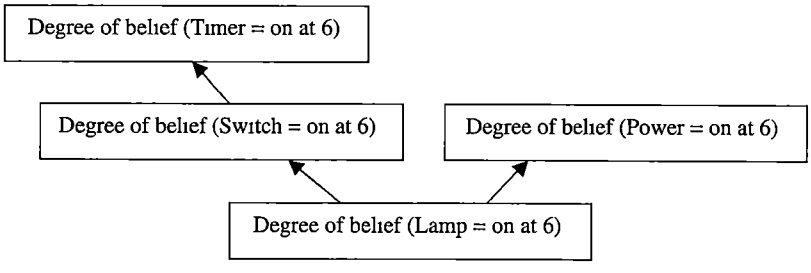


Figure 4 The Causal Bayes Net of Reasoning from Perception

If reasons are causes, then the causal structure for reasoning from perception is exactly reversed from the causal structure for reasoning to a forecast If Degree of belief(X) $\rightarrow$ Degree of belief(Y) in the Causal Bayes Net of Reasoning to a Forecast, then Degree of belief(X) $\leftarrow$ Degree of belief(Y) in the Causal Bayes Net of Reasoning from Perception The reversal of edges, however, means that the two causal Bayes nets—one of forecasting and the other of perception—do not agree in the constraints they impose on the joint probability distribution for degrees of belief In this example, the Markov property applied to the Causal Bayes Net of Reasoning to a Forecast, implies that Degree of belief(Power = on) = x is unconditionally independent of Degree of belief(Switch = on) = y, for all x, y The Causal Bayes Net of Reasoning from a Perception implies no such thing We are whipsawed

III COMBINING THE BAYES NETS

Human perception, we think, is often in part top-down, driven by prior conceptual structure and prior degrees of belief For a Bayesian agent whose reasons are causes, the problems just discussed suggest that perception that accords or conflicts with a prior degree of belief should have a top-down contribution In order for sensation to cause our imagined rational agent to form a new degree of belief that the lamp is lit, and to do so in a way that allows a Bayesian updating of the value of the agent's degrees of belief that the switch is on and that the power is on, the new degree of belief that the lamp is lit must be the collaborative, interactive effect of the values of her degree of belief that the switch is on, of her degree of belief that the power is on, and of the sensory input The sensor input does not *itself* change the value of Degree of belief(Lamp = on), but rather it changes the degree of belief that the lamp is lit *given the values of parents of Degree of belief(Lamp = on)* in the Causal Bayes Net of Reasoning to a Forecast For different values of Degree of belief(Power = on) and Degree of behef(Switch = on), the input of sensation will result in different values of Degree of belief(Lamp = on), and so the intervention of sensation will not make the agent's Degree of belief that the Lamp is on independent of her other Degrees of belief

Since reasoning goes from beliefs about circumstances to forecasts of perceptions, and from perceptual changes in belief to new beliefs about circumstances, it seems that the "reasons are causes" view requires a representation of the causal connections that likewise goes in both directions It seems that we need, in other words, a cyclic causal graph among degrees of belief, with appropriate associated probabilities

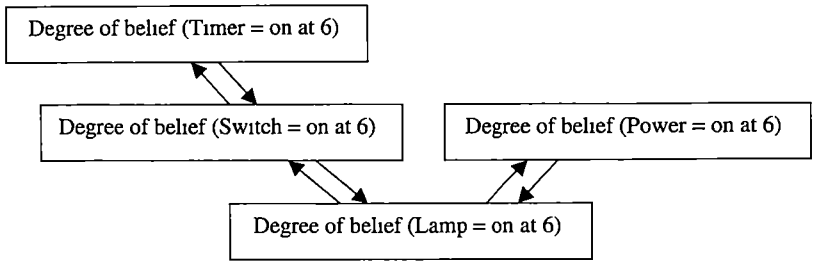


Figure 5 The Combined Causal Bayes Net

The problem is we do not know much about cyclic graphical representations of causal relations, or how to update them by Bayes's rule, and what we do know is problematic for this view In the scientific

literature, causal Bayes nets are generally taken to be acyclic, but that is not strictly necessary One can have networks with cycles, even with cycles that have edges in each direction between two variables Probabilistic constraints that generalize the Markov property for acyclic networks still hold, necessarily, for linear cyclic systems, and can consistently be assumed for cyclic networks with variables that have a finite range of possible values So we might consider whether the sensory input can nudge the degrees of belief in values of a variable, which nudges the degrees of belief in its parents, which nudges the degrees of belief in the variable again, which nudges and so on, until an equilibrium is reached

That is certainly possible, but there are two related difficulties How can updating on evidence occur, and can it be Bayesian? Consider the second difficulty first, in the simplest case in which the variable that is directly influenced by sensory inputs, denote it by S, has a single parent variable, Y The idea is that the value of S causes the value of Y to be updated, which causes the value of S to be updated, and so on, until no more changes result On the Bayesian perspective, each step in each direction, no matter how implemented in the brain, should result in updating one of the variables conditional on the currently updated value of the other, and we should therefore expect that at equilibrium the joint degree of belief in S and Y together should be the product of their conditional probabilities on each other for all values of S and Y, $\mathrm{DOB}(S,Y) = \mathrm{DOB}(S \mid Y)\mathrm{DOB}(Y \mid S)$ But this equation implies that Y and S are independent[7] Applied to our example, upon learning that the lamp is not lit at 6, the agent's degrees of belief would then be altered in such a way that the degree of belief that the switch is on and the degree of belief that the timer is on have no relation to one another We should not welcome such a theory

Not only does a Bayes Rule requirement for updating lead to absurd results in cyclic networks, no correct updating algorithm is known for such systems and certainly no algorithm of the kind that neural systems plausibly implement for acyclic Bayes nets[8] Some other resolution is needed

IV DYNAMICS TO THE RESCUE?

The general problem is that the causal direction of influences of degrees of belief must go one way when forecasting, and the reverse

[7] $Pr(S \& Y) = Pr(S \mid Y)Pr(Y \mid S) = Pr(S \& Y)Pr(Y \& S) / Pr(Y)Pr(S) \Rightarrow Pr(Y)Pr(S) = Pr(Y \& S)$

[8] However, procedures for computing correlations resulting from interventions on a single variable in linear cyclic systems have been known for a long time

direction when learning from experience, and the conditional probabilities of the changes must be in phase If we can presume that forecasting and learning are not simultaneous, then there is a Bayesian solution, using structures that are sometimes called "dynamical Bayes nets" but which are really the same sorts of structures we have considered so far, except that the variables—in this case degrees of belief—are indexed by time Consider the structure

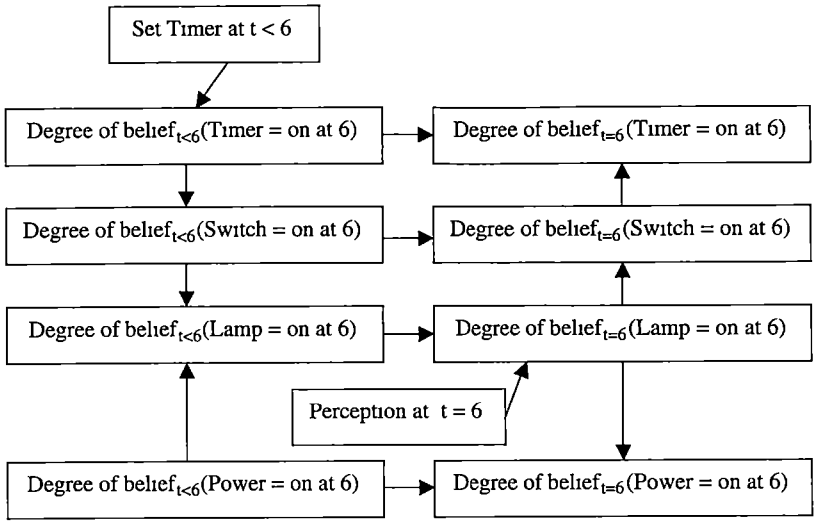

Set Timer at t < 6
Degree of belief$_{t<6}$(Timer = on at 6)
Degree of belief$_{t=6}$(Timer = on at 6)
Degree of belief$_{t<6}$(Switch = on at 6)
Degree of belief$_{t=6}$(Switch = on at 6)
Degree of belief$_{t<6}$(Lamp = on at 6)
Degree of belief$_{t=6}$(Lamp = on at 6)
Perception at t = 6
Degree of belief$_{t<6}$(Power = on at 6)
Degree of belief$_{t=6}$(Power = on at 6)

Figure 6 The Dynamical Bayes Net of Reasoning to Forecasts and from Perception

All of the probabilities of conditional degrees of belief in this network can be consistently estimated by Bayes's Rule, even while the values of the variables—the degrees of belief—are themselves determined, up to chance variation, by Bayes's Rule applied to the external evidence, setting the timer and sensation [9]

So we have a solution in which reasoning is Bayesian almost all the way through, and reasons are causes We do not know of any neural implementation of dynamic Bayes nets, and any neural realization that involves both forecasting and learning, rather than only visual recognition, will not be localized in the visual cortex Verifying the hypothesis that, in humans, reasoning is Bayesian and reasons are

[9] Strictly, this implements the representation of interventions proposed in Spirtes, et al (*op cit*), rather than the representation in Pearl (*op cit*)

causes would be difficult, but at least we have some idea of the conditions that need to be satisfied

However, this is in some respects an odd and incomplete solution As in our story, agents must be able to reason hypothetically, for otherwise we could not rationally plan But hypothetical reasoning requires hypothetical degrees of belief the expectations one would have if one believed some set of assumptions, be they assumptions about causal structure or assumptions about values of variables ascribed, hypothetically, to the world, or assumptions of some other kinds When reasoning under a hypothesis, hypothetical degrees of belief, or their changes, ought still to be causes of other hypothetical degrees of belief in something like the way we have described, at least if the Bayesian, "reasons are causes," account is correct There is no difficulty in this from the point of view of an omnipotent Bayesian calculator (the kind philosophy generally assumes), and there is no difficulty in principle from the point of view of a programming system, provided the number of alternative hypotheses is not too large the degrees of belief on each hypothesis are computed, and averaged with weighting by the degrees of belief in the various hypotheses But how a neural system could implement such reasoning, distinguishing between the hypothetical and the all-things-considered nonhypothetical degrees of belief, remains to be discovered or not And that is not the only inadequacy

The agent must have just the right course of reasoning instantiating just the right Dynamic Bayes Net, and that must somehow, mysteriously, be determined by the agent's goals and epistemic circumstances at the moment For example, if the agent is not home at 6 00 P M to perceive the state of the lamp, she may follow the forecast to 6 00 P M with another forecast from 6 00 P M If instead of seeing the state of the lamp, she sees the power is off, she will have a quite different sequence of changes of belief The correct course of reasoning must somehow be generated on the fly, and we have no account of how that is done The machine still has its ghost

CLARK GLYMOUR
DAVID DANKS

Carnegie Mellon University and
Florida Institute for Human and Machine Cognition

COMMENTS AND CRITICISM

MORAL RESPONSIBILITY WITHOUT ALTERNATIVE POSSIBILITIES?*

The article I am commenting on, "Libertarianism and the Philosophical Significance of Frankfurt Scenarios," represents a significant turn in David Widerker's thinking about moral responsibility, or, more precisely, moral blameworthiness [1] He now holds that an agent can be morally blameworthy for an action of his even if it was not within his power to avoid performing it This "avoidability-independent account of moral blame," as Widerker calls it, contrasts with his previous defense of avoidability as a requirement of moral responsibility The article contains deep insights and a lot of good argument My purpose in this comment, however, is mainly critical After a brief presentation of Widerker's new position, I will state why I do not find it convincing

In an important and influential article,[2] Widerker put forward a powerful defense of the Principle of Alternative Possibilities (PAP) against criticisms based on so-called Frankfurt-style examples or Frankfurt scenarios According to PAP, in Widerker's recent formulation, "[a]n agent *S* is morally blameworthy for performing a given act *V* only if he could have avoided performing it" (163) [3] Frankfurt scenarios are cases in which an agent, fully on his own, does something which he would have been caused to do anyway by an alien factor had he shown some sign that he was not going to do it on his own

*This paper falls under the research project "Belief, Responsibility, and Action" (HUM2006-04907), awarded and supported by the Spanish Ministry of Education and Science, to which thanks are due Let me also thank Milagro Garrido, Tobies Grimaltos, Javi Hernandez, Marta Moreno, Eduardo Ortiz, Sergi Rosell, and Lino San Juan for useful comments on previous versions of this paper

[1] Widerker, "Libertarianism and the Philosophical Significance of Frankfurt Scenarios," this JOURNAL, CIII, 4 (April 2006) 163–87

[2] Widerker, "Libertarianism and Frankfurt's Attack on the Principle of Alternate Possibilities," *Philosophical Review*, CIV, 2 (1995) 247–61

[3] Harry Frankfurt's formulation of PAP was as follows "A person is morally responsible for what he has done only if he could have done otherwise" ("Alternate Possibilities and Moral Responsibility," this JOURNAL, LXVI, 23 (December 4, 1969) 829–39, on p 829) Widerker's formulation is more general, in that an alternative to performing a given action can be just to avoid performing it, and not necessarily to perform a different action It is plausible to think that this was the spirit of Frankfurt's formulation as well

0022-362X/07/0409/475–86 © 2007 The Journal of Philosophy, Inc

If cases of this sort are to imply the falsity of PAP, they have to satisfy at least two conditions First, the circumstances that make the act unavoidable for the agent should not cause the agent to perform it Widerker calls this condition *IRR*.[4] Here is Widerker's own formulation "(IRR) There may be circumstances that in no way bring it about that a person performs a certain action, nevertheless, those very circumstances make it impossible for him to avoid performing that action" (164) A second condition is, of course, that the agent in such scenarios is morally responsible for his act If at least one Frankfurt-style example meets these conditions, we have a case in which an agent is morally responsible for something he does even if he could not have avoided doing it, and PAP is then false [5] A rationale for condition IRR is that, if the act is unavoidable owing to circumstances that causally necessitate it, then the judgment that the agent is morally responsible for his act becomes unstable, and satisfaction of the second condition is jeopardized Indeed, if that is the case, incompatibilists will (understandably enough) refuse to accept the agent's moral responsibility But even compatibilists may hesitate for, depending on the nature of the circumstances that causally necessitate the action, it may appear as forced, coerced, or otherwise not freely performed

Widerker contended that Frankfurt scenarios do not satisfy the IRR condition for choices or decisions, which are the natural *loci* of moral responsibility for libertarians His argument for this contention took the form of a dilemma [6] For simplicity, suppose that the alien, inactive, though lurking factor in Frankfurt scenarios is an agent, Black Think of the sign that Black employs for *not* intervening in the agent's (call him Jones) spontaneous process of deliberation and decision Now, this is the dilemma Either this sign is causally sufficient for Jones's decision, or it is not If it is, then condition IRR is not satisfied, for the decision is made unavoidable by circumstances that causally determine it If it is not, so that the sign is only a more or less reliable indicator of Jones's future decision, then there is no reason to think that the decision is unavoidable, and condition IRR is violated again In either case, PAP is not refuted On the first horn, there are no alternatives for Jones's deci-

[4] I have not been able to ascertain why he calls it this

[5] In my recent book *Moral Responsibility The Ways of Scepticism* (New York Routledge, 2006), chapter 2, I have extensively defended PAP against Frankfurt-inspired and other criticisms of it In the present paper I defend PAP further against Widerker's criticism

[6] Cf Widerker, "Libertarianism and Frankfurt's Attack on the Principle of Alternate Possibilities," pp 251–56

sion, but he is not (for incompatibilists, at least) morally responsible for it On the second horn, Jones may be morally responsible for his decision, but it certainly seems that he had alternatives to it he could have decided otherwise, or at least could have avoided the decision he actually made [7]

As Widerker writes "The worry behind the dilemma objection is understandable If in an IRR-situation the agent's decision is not causally determined, then why is it not within the agent's power to refrain from it?" (166) Since Widerker's forceful presentation, the dilemma defense has exerted strong pressure on the construction of prima facie plausible Frankfurt scenarios Widerker holds that none of the attempts to design such scenarios in response to the dilemma defense has been successful However, unlike what he used to think, he now contends that IRR-situations are nonetheless possible "In the past, I have been skeptical about the prospects of providing a successful example of an IRR-situation Since then, however, I have come to believe that such situations *are* conceptually possible" (166) Let us see the way in which Widerker has tried to show the possibility of these situations

A chief difficulty in conceiving of IRR-situations is that it does not seem possible to reconcile the fact that the decision (or action) is not causally determined with its unavoidability In an indeterministic setting, alternative pathways of some sort are bound to be present, and in some of these alternative pathways the decision is not made or the action is not performed In fact, Widerker does not actually solve *this* difficulty As we will see, his purported examples of IRR-situations are not such that in them the agent's decision is strictly unavoidable However, Widerker resorts to a thesis that has emerged from the discussion about Frankfurt scenarios, namely, that it is not strictly necessary, for a situation to count as a counterexample to PAP, that in it the agent's decision is actually unavoidable It is enough that the alternative pathways that may be present in it, which may include the agent's not making the decision he actually makes, are not within the agent's power or control, it is enough, in Widerker's terms, that they are not "actionally accessible" to the agent "Call a state of affairs (event) *E actionally accessible* for an agent *S*, relative to a time *T*, if and only if *S* has it within his power at *T* to bring about *E*" (167)

[7] A related defense of PAP was put forward by Robert Kane in his *Free Will and Values* (Albany SUNY, 1985) However, it was Widerker's forceful presentation of it that brought this strategy to the center of the discussion about PAP and Frankfurt-style objections to it

Alternatives that are only causally possible,[8] but not actionally accessible, to an agent do not suffice to defend PAP against counterexamples To see why, imagine that, in a Frankfurt scenario, an agent deliberates competently and, on the basis of his own reasons, makes a morally wrong decision Suppose that this decision is avoidable *only* as follows before making it, the agent might, inadvertently and involuntarily, have stumbled on a stone and got hurt, with the result that he would have stopped deliberating and not have made the decision he actually made If, in this example, the agent is morally responsible for his decision, it does not seem that such an alternative has anything to do with his responsibility this alternative is not morally significant Actional accessibility is a necessary condition for an alternative to be morally significant or, to use John Martin Fischer's well-known term, to be *robust* enough to explain the agent's moral responsibility[9] So, actionally inaccessible alternatives cannot be legitimately pointed to in order to vindicate PAP against putative counterexamples

On the basis of the preceding remarks, we can understand Widerker's strategy It consists in constructing examples where the agent's decision is not causally determined, but where the alternatives that are left, though causally possible, are not actionally accessible to the agent (and so not morally significant or robust) Strictly speaking, then, these are not properly IRR-situations, for the agent's decision is not unavoidable They are, however, examples of a more general notion that includes the original IRR-situations as particular cases of it Widerker calls this more general notion or condition *G-IRR*. "There may be circumstances that in no way bring it about that a person performs a certain action, nevertheless those very circumstances leave him with no morally significant alternative to performing that action" (175) Notice that having no alternative to what one does is a particular way of having no morally significant alternative to it Corresponding to this more general definition of an IRR-situation, Widerker suggests the following more general formulation of PAP "An agent *S* is morally blameworthy for *V*-ing at *T* only if *S* had a *morally significant alternative* to *V*-ing at *T*" (175, note 24) It is plausible to think that this formulation is faithful to the

[8] Widerker defines a causally possible state of affairs *E*, relative to a time *T*, as follows "*E* obtains in some possible world *W* that up until (and including) *T* has the same history and the same laws of nature as the actual world" (167)

[9] It was Fischer, in fact, who first pressed this point, according to which alternative possibilities in Frankfurt scenarios have to be robust, in the sense of being explanatorily relevant for the agent's moral responsibility, if resort to them is to be successful as a defense of PAP Cf Fischer, *The Metaphysics of Free Will* (Cambridge Blackwell, 1994), pp 140–47

original inspiration of PAP alternative possibilities mentioned in this principle are plausibly taken to be morally significant or robust Strictly speaking, then, the scenarios Widerker depicts are examples of G-IRR-situations, rather than of IRR-situations, and they are purported counterexamples to the general formulation of PAP, rather than to the original formulation of it Having this in mind, however, we shall use "IRR-situations" to refer to G-IRR-situations (which, of course, include IRR-situations proper), and with "PAP" we will refer to the general formulation of this principle we have just indicated Let us now have a look at Widerker's purported examples of IRR-situations

Widerker depicts his IRR-situations on the basis of an initial example of a Frankfurt scenario, which he dubs "Promise Breaking" The example features Jones, who has promised his uncle to visit him in the hospital before an important operation that the old man will undergo early next morning On his way to the hospital, Jones meets Mary, a former girlfriend of his, whom he has not seen for a long time Mary tells Jones that she will be in town for just a couple of hours and that she would most like to spend those hours with him Jones feels strongly attracted by Mary's proposal He knows, however, that, if he accepts it, he will not be able to keep the promise he made to his uncle, for the visit time will be over So he deliberates whether to keep his promise Unbeknownst to him, however, another person, Black, wants Jones to break that promise and has the means to ensure this result However, Black will remain fully inactive unless Jones shows a sign that he will finally decide to keep his promise In fact, Black has no need to intervene, for Jones shows no such sign and decides on his own to break his promise and to spend time with Mary (cf 164)

Widerker contends that this example falls prey to the dilemma defense and is not an IRR-situation However, elaborating on this example, Widerker constructs two putative examples of IRR-situations The first he calls "Z-Persons" Suppose that there actually are persons (Z-Persons) who, in morally laden choice situations in which they feel a strong temptation to act immorally, invariably behave as follows either they succumb to the temptation and make the wrong decision or they lose consciousness, regaining it shortly afterwards Widerker writes

> Now suppose that Jones is such a person, and he is in the kind of situation described in Promise Breaking, except that this time there is no Black waiting in the wings Suppose also that Jones does not know that he is subject to the said behavioral pattern As things transpire, Jones decides at time T to break the promise on his own (169)

Widerker contends that this is an IRR-situation Jones's decision to break his promise is not causally determined It might not take place However, the only alternative to it in which Jones does not make that decision, namely Jones's loss of consciousness (assuming that this is not under Jones's control), is not actionally accessible to him, and so is not morally significant (or robust) Jones, then, does not have it within his power to decide to keep his promise, which would be a morally significant alternative Notice also that the example does not succumb to the dilemma defense, for it does not feature a counterfactual intervener or a sign of Jones's future decision concerning which the dilemma could arise

Widerker's second example is called "Brain-Malfunction-W" It again features Jones's deciding at time T to break his promise A different decision (which would be a morally significant alternative) is not open to him, for

> shortly before beginning to deliberate, he undergoes a neurological change as a result of which one of the (neurological) causally necessary conditions for his deciding otherwise, a condition which we may call N, does not obtain Let us also assume that all this is unknown to Jones (who believes that he can decide to keep his promise), and that N's absence does not affect his deliberation process (170)

Since N is causally necessary for Jones's decision to keep his promise, N's absence is causally sufficient for Jones's *not* deciding to keep his promise This absence, however, is not causally sufficient for his (actual) decision *not* to keep his promise This decision might not take place, but the alternatives to it (for instance, becoming distracted and temporarily forgetting about his deliberation), though causally possible, are not actionally accessible to Jones, and so not morally significant

It seems, then, that we have two examples of an IRR-situation More exactly, they are examples of a G-IRR-situation, for, though Jones's decision is not, strictly speaking, unavoidable, the alternatives to it are not morally significant However, as Widerker himself accepts, in addition to its being an IRR-situation, an example has to satisfy a second condition if it is to be a counterexample to PAP, namely that the agent is morally responsible for what he did In the purported examples of IRR-situations featuring Jones, he is to be morally blameworthy for his decision to break his promise, for otherwise PAP will remain safe Now, provisionally leaving aside the question whether Widerker's examples are in fact IRR-situations, my criticism of Widerker's position will rest crucially on the contention that his examples do not satisfy the indicated condition of moral blameworthiness More cautiously, my contention would be that there are fairly

strong reasons against the judgment that, in the proposed examples, the agent is morally blameworthy for the decision he makes

It is quite clear that Widerker believes that, in both of his examples, Jones is morally blameworthy for breaking his promise Widerker's new, avoidability-independent account of moral blame (AI-BL) takes the following form

(AI-BL) An agent S is morally blameworthy for deciding-to-V at time T if and only if
- (i) S decides-to-V at T on his own
- (ii) S decides-to-V at T being aware that doing so is morally wrong
- (iii) S believes that he could have decided otherwise, or believes that he has no good reason to think that he could not have decided otherwise
- (iv) S does not think that, in the circumstances, deciding-not-to-V at T is irrational (180)

It seems that Jones, both in Z-Persons and in Brain-Malfunction-W, satisfies these four conditions, so, if the account is correct, he is morally blameworthy for breaking his promise The account does not require that the agent be actually able to decide otherwise, but only that he believe he can do so, and this is something that Jones believes in both cases, though his belief is false

As I said earlier, I think there are serious reasons for doubting that Jones is actually morally blameworthy in either case Let me place my criticism in the context of contemporary dialectics concerning moral responsibility My objections to Widerker are related to what Fischer has called "weak reasons-responsiveness," which he takes to be a requirement for moral responsibility [10] Roughly speaking, in deliberating and deciding, an agent satisfies weak reasons-responsiveness just in case, keeping constant the mechanism of deliberation and decision-making that he employs in the situation at hand, there are some possible scenarios or possible worlds in which there is a sufficient reason to decide and do otherwise, he recognizes that reason and decides and does otherwise It is very plausible that something like this condition is necessary for moral responsibility Think of someone (the example is Fischer's) who decides to steal a book and does so Fischer writes "If (given the operation of the actual kind of mechanism) he would persist in stealing the book even if he knew that by so acting

[10] Cf Fischer, *The Metaphysics of Free Will*, pp 166–67 Fischer has slightly modified the requirement in his book *Moral Responsibility and Control* (New York Cambridge, 1998), coauthored by Mark Ravizza In this book, Fischer and Ravizza call the modified requirement "moderate reasons-responsiveness " I will ignore the distinction, for it is not essential to my argument

he would cause himself and his family to be killed, then the actual mechanism would seem to be inconsistent with holding him morally responsible for his action."[11] This thief's mechanism of deliberation and decision-making is not responsive to reasons, it is a faulty mechanism, which compromises the agent's capacity for rational deliberation and decision-making and undermines the judgment that he is morally blameworthy for his decision and action. Whether a certain process satisfies Fischer's weak reasons-responsiveness depends on the truth of certain counterfactuals about what the agent would decide and do if certain conditions were to hold, but it does not require that the agent has access to the possible worlds where these conditions hold. It does not assume, then, that moral responsibility for a decision or action requires that an agent is actually able to decide and do otherwise.

Now, a related condition should hold for the specific case of moral reasons. Some degree of responsiveness to moral reasons is plausibly taken to be a requirement for moral responsibility. If, keeping constant the mechanism of deliberation and decision-making that an agent employs in a particular situation, and supposing that he faces a temptation to act immorally, he would never decide in a morally right way, no matter how strong and decisive his moral reasons may appear to him, it seems that something is wrong with the agent's deliberative faculties in this situation. And this stands in the way of judging him morally blameworthy for his decision. Now, this seems to be precisely the case with Jones in Widerker's examples. Let me argue for this point.

As Fischer's weak reasons-responsiveness suggests, an important assumption of the judgment that an agent is morally responsible for what he decides is, for compatibilists and incompatibilists alike, that the decision arises out of a normal process of practical deliberation and decision, which in turn presupposes in the agent a sound capacity for practical and moral reasoning. This capacity can be impaired in several ways. One of these ways may be the belief that one does not have alternative decisions or actions within one's power. This belief may temporarily damage the indicated capacity, vitiate the process of deliberation, and affect one's moral responsibility. In Widerker's account, the third condition rules out this possibility, and we are entitled to assume that, in both examples, Jones believes (falsely, as it happens) that he can decide to keep his promise. But there are other ways in which the impairment may occur. Being insensitive to moral reasons (being "morally blind") is one of them. Widerker's second condition seems to rule out this possibility as well. In his examples,

[11] Fischer, *The Metaphysics of Free Will*, p. 167.

we should assume, it seems, that Jones is sensitive to moral reasons for keeping his promise But, beyond being sensitive to moral reasons, an agent with a normal, unimpaired capacity for moral and practical reasoning and decision-making should also be able *to decide* (and possibly act) according to these reasons And this is an ability that Jones lacks in both of Widerker's examples As Widerker depicts the cases, the actual apparatus of practical reasoning and decision-making that Jones employs is such that he is not able to decide as moral reasons counsel Deciding to keep his promise is not causally possible for Jones, and so is not within his power And if this is the case, Jones's capacity for moral deliberation and decision-making is seriously impaired, so that the judgment that he is morally blameworthy for the decision he makes in the circumstances becomes highly dubious, to say the least Let me clarify and develop this point

First of all, I would like to emphasize that my objection to the judgment that Jones is morally blameworthy for his decision to break his promise does not rest on the fact that he lacks morally significant alternative possibilities, for this would clearly beg the very question at issue My objection has rather to do with the fact that the apparatus of deliberation and decision-making that Jones is using in the actual sequence is abnormal or impaired In the context of my objection, the fact that Jones cannot decide to keep his promise, no matter how strong or decisive the moral reasons for this decision may appear to him, is only a symptom or a consequence of the faulty nature of his actual apparatus of practical and moral reasoning And it is this faulty nature, and so a feature of the actual sequence, rather than the lack of robust alternatives, that backs my doubt about Jones's blameworthiness

The sense in which Jones's actual apparatus of practical and moral reasoning is faulty can be seen as follows Consider first Z-Persons Imagine that, on his way to the hospital, Jones meets a friend of his who is a doctor Jones tells his friend about the operation that his uncle is about to undergo and he comes to know from his friend how dangerous and risky the operation is, and how likely it is that his uncle will not survive it We may also suppose, if we want, that the old man has generously helped Jones in many ways during his life, and that Jones feels deeply grateful to him for that Jones now sees in an especially clear way what strong and decisive moral reasons he has for visiting his uncle Afterwards, he meets Mary as in the original story and, though he still feels strongly tempted by the prospect of spending time with her, he now sees clearly that his reasons for visiting his uncle are more important and decisive than his reasons for staying with Mary However, as he is a Z-person, he is, in fact,

causally unable to do what a normal moral agent would rightly be expected to do in the circumstances, namely, deciding to keep his promise and visiting his uncle Now, I would think that an agent who, in a deliberative situation like this, is causally unable to make the right moral decision, does not have a sound and unimpaired capacity for practical reasoning and should not be judged to be (at any rate fully) morally blameworthy for the decision he makes, no matter whether one is a compatibilist or an incompatibilist

Moreover, it would be pertinent to ask for more information about the example, for, depending on how it is interpreted, there might be additional reasons for criticism In particular, it would be important to know under what circumstances the nonrobust alternative of Jones's fainting is supposed to take place Suppose that it may occur at any time while Jones is deliberating and merely considering the reasons for either decision In this case, it seems that he does not have enough control over his decision at T to break his promise Though he actually made that decision, he might equally have fainted instead, so that the judgment that he is morally blameworthy for it is eroded On this interpretation, Widerker's example is not immune to traditional luck objections to incompatibilism Suppose instead that Jones would faint only as a result of his attending seriously to his moral reasons and his making an effort to decide according to them In this case, Jones may be morally blameworthy for his decision to break his promise, but now he has morally significant, robust alternatives Jones would therefore be blameworthy for his decision partly because he should and could have made an effort to decide otherwise but did not On this interpretation, the example would not be a G-IRR-situation

Similar remarks can be made with respect to Brain-Malfunction-W In this example, after neurological condition N has ceased to obtain, Jones will not be able to decide according to his moral reasons, no matter how strongly and decisively these reasons may appear to him But this causal impossibility of alignment between reasons and decision is a clear sign that something is going wrong with Jones's deliberative and decision-making apparatus in that situation It seems, then, that Jones does not reach his decision to break his promise with a sound and unimpaired capacity for practical and moral reasoning, and this undermines the justification for judging him morally blameworthy for his decision On the other hand, depending on what the absence of condition N allowed Jones to do, the example might not be a G-IRR situation By hypothesis, this absence allows Jones to appreciate the force of moral reasons Suppose that it also allowed him to adopt a favorable attitude toward complying with them, as a preliminary step in the process of deciding to keep his promise Even if

the process could not finally be completed, adopting such an attitude would seem to be a morally significant alternative, and the example would not be a G-IRR-situation Suppose, however, that the absence of *N* did not even allow him to adopt a favorable attitude toward a decision that would honor moral reasons, or to take any other preliminary step addressed to that decision If so, this would reinforce our initial suspicion that, in the situation described in the example, we were not facing an agent with unimpaired faculties for practical deliberation An agent unable to take any steps towards deciding according to moral reasons, no matter how strong or convincing he may find them, does not seem to be a normal deliberator, and the case for not holding him (fully) morally responsible is enhanced [12]

Though the possibility that Widerker, or someone else, can come up with other cases of IRR-situations in which the agent is morally responsible for his decision cannot be ruled out a priori, there are some general considerations, prompted by the preceding criticisms, that suggest that such a possibility is not in the offing and even that cases like it may not be possible To conclude, let me argue for this point

An important advantage of Frankfurt scenarios is that they tend to raise a natural and very strong intuition that, in them, the agent is morally responsible for his decision, for this decision is, at least apparently, ensured only by the presence of a purely counterfactual factor, which remains fully inactive throughout the whole process The agent seems to deliberate and decide spontaneously, with no apparent hindrance or constraint Moreover, there is nothing that prevents us from assuming that the agent's actual apparatus of practical reasoning is perfectly sound and reasons-responsive Suppose, nonetheless, that Frankfurt scenarios fall prey to the dilemma objection or are otherwise irredeemably flawed [13] In Widerker's alternative proposal, however, the intended result, namely either a morally wrong decision or morally irrelevant alternatives, is not ensured by a counterfactual factor (which would open the door to the dilemma

[12] Though I will not develop this point here, I am confident that similar criticisms of Widerker's proposal can also be raised on the basis of other views of control and moral responsibility, such as Frankfurt's, Gary Watson's, or Susan Wolf's It seems clear that, in Widerker's examples, and under quite plausible assumptions, Jones will not be able to decide and act in accord with either his second-order volitions or his values

[13] Given the number, variety, and ingenuity of Frankfurt-friendly examples now on offer, I am aware that this is a bold assumption indeed, but let us accept it for the sake of argument Some support for it can be found in my *Moral Responsibility* (chapter 2) There I try to show that, even if some of the recently proposed examples can actually escape the dilemma objection, they may face other, no less significant, difficulties

NOTES AND NEWS

Es nehmet aber / Und giebt Gedachtniß die See

The editors are deeply saddened to report that Michael Frede has died He drowned on August 11th, 2007, near Delphi, while swimming in a small cove off the Gulf of Corinth Professor Frede was 67 years old

Frede was born in Berlin on May 31st, 1940 In 1943 his family moved from Berlin, first to Lippstadt (Westphalia), then, in the early 1950s, to Hamburg, there he attended the Sankt Ansgar-Gymnasium (a Jesuit school), receiving his *Abitur* in 1959 Frede studied at the Universities of Munich (1959–60), Hamburg (1960–62), Oxford (1962–63), and Göttingen (1963–66), where he was awarded his doctorate (in 1966) for a dissertation on Plato's *Sophist*, written under Günther Patzig's supervision and published as *Prädikation und Existenzaussage* (1967) In 1972 Frede's study of Stoic logic was accepted as his *Habilitationsschrift* by the Faculty of Philosophy at Göttingen, it was published as *Die Stoische Logik* (1974) While Frede was educated primarily in Germany (though G E L Owen, with whom he studied in Oxford, remained an important influence), he spent the bulk of his academic career in the English-speaking world After working for a few years as an *Assistent* in Göttingen, he came to the United States, first to Berkeley (1971–76), then to Princeton (1976–91), in 1991 he left the United States for England, where he held the chair of the History of Philosophy at Oxford until retiring in 2005 Following his retirement from Oxford, Frede moved to Athens

His *Essays in Ancient Philosophy* (1987) gathers most of the articles written before 1986 These essays already testify to the strikingly wide range of authors and topics investigated by Frede, they include papers, many by now classics, on Aristotle's *Categories* and *Metaphysics*, the Stoic conception of causation and its antecedents, Stoic epistemology and psychology, epistemological debates in ancient medicine, the relation between ancient works on grammar and philosophical, especially Stoic, discussions of logic and language, they also include two groundbreaking treatments of Academic and Pyrrhonian Scepticism, which were later again reprinted, in a volume co-edited with M Burnyeat *The Original Sceptics A Controversy* (1997) For the academic year 1984-85, Frede returned to the city of his birth, as a Fellow at the *Wissenschaftskolleg zu Berlin*, there he and G Patzig co-authored their magisterial *Aristoteles 'Metaphysik Z' Text, Übersetzung und Kommentar* (1988), a work that was immediately recognized as a towering achievement in Aristotelian studies, not least on account of its defense, at once subtle and vigorous, of the controversial claim that, for Aristotle, the forms of particular substances are themselves individuals in their own right

The years in Oxford brought further collaborative projects A *Festschrift* for Günther Patzig, co-edited with G Striker *Rationality in Greek Thought* (1996) A volume, co-edited with P Athanassiadi, exploring the relationship between Christian and pagan conceptions of monotheism *Pagan Monotheism in Late Antiquity* (1999) And the proceedings of the 14th Symposium Aristotelicum, co-edited with D Charles *Aristotle's* Metaphysics *Lambda* (2000) Frede furnished each of these collections not only with a major paper of his own, but also with a substantial introduction (in one instance, co-authored) Indeed, his lengthy introduction to the Aristotle volume and his contribution to the symposium (an essay on Lambda 1) form what could virtually stand on its own as a short, incisive monograph on both this book of the *Metaphysics* as well as the larger metaphysical project Aristotle is pursuing

In the fall of 1997 Frede returned to Berkeley as a visitor, to give the Sather Classical Lectures The six lectures were devoted to the origins of the notion of free will—and thus, the origins of the relevant notions of a will and of freedom—matters that had long been of interest to him and that continued to occupy him in subsequent years Until

and unless those lectures are published, his paper, "John of Damascus on Human Action, the Will, and Human Freedom" (in K. Ierodiakonou, ed., *Byzantine Philosophy and its Ancient Sources* (2002)), will have to serve as the record of his deep knowledge of, and careful reflection on, this cluster of topics Frede explores and explains the picture of the will (and its freedom) which the late seventh-/early eighth-century theologian John of Damascus—the last figure from the Greek East regarded as an authority in the Latin West—presents in several of his works Frede shows how John of Damascus draws on ideas that ultimately derive from Aristotle and from the Stoics, as mediated by Neoplatonism and various Church Fathers In tracing out and seeking to clarify the different elements in John of Damascus's picture, Frede also sheds much light on those earlier discussions of the will, freedom, and related notions His account of how John of Damascus influenced various philosophers in the Latin West, including Robert Grosseteste and Thomas Aquinas, reveals a side of Frede's scholarship that surfaces only now and then in the published writings but was often on display in his teaching his extensive knowledge of medieval philosophy

If philology is, as it is sometimes said to be, the art of reading slowly, then Frede was, in his seminars, a consummate practitioner of this art Indeed, to the casual observer, it could seem as if he made progress at only a glacial pace, if at all he thought about *every* element of the assigned text—beginning with its title—in the most fine-grained way, he laid out and explored, in exquisite detail, a whole range of interpretative options, and above all, he sought to illuminate philosophically each word, sentence, or passage under discussion by examining it in the light of other passages from the chosen text, further works by the same author, or still other texts, often ones which at first glance might seem far removed from his starting point And if Frede had, by the end of the term, moved only a few pages into the chosen text, he had given to the seminar's participants, as abundant recompense, a rich and detailed framework for thinking productively about that text and the issues it raised

The published work endures as a valuable, and valued, resource that others will continue to be able to draw on and engage with While Frede always was, throughout his career, deeply committed to the study of Plato and Aristotle—the "classical" philosophers of classical antiquity—his helping to expand greatly the field of inquiry, in two directions, constitutes an equally important part of his legacy He was at the forefront of those exploring Hellenistic philosophy and, especially during the last two decades, the philosophy of later antiquity Stoicism, Scepticism, and Neoplatonism now can figure even in undergraduate classes in a way they simply could not in earlier years Perhaps still more significantly, Frede (along with others) turned to the large number of grammatical, medical, and theological writings from antiquity, and in doing so, demonstrated the importance of going rather far beyond the confines of overtly "philosophical" writings for the project of attaining a systematic understanding of ancient philosophy

His erudition and philosophical acumen were immense, as even a casual reading of the written work reveals Yet what was perhaps most impressive was his astonishing *command* of a vast range of texts, his extraordinary ability to recall even the minutest of details The authors and issues Frede worked on were difficult, in many cases made more difficult by the problematic nature of the texts transmitted to us His scholarship was thorough, painstaking, and nuanced, often intricate, and at times complex in its particulars But this scholarly activity was guided by a simple and wholly straightforward conviction that the rigorous study of the history of philosophy was a worthwhile and intellectually rewarding enterprise, a conviction only strengthened by the results he and many of his colleagues achieved And in engaging in that enterprise, Michael Frede was above all (as he himself might have said) trying to think clearly

UNITED STATES POSTAL SERVICE

Statement of Ownership, Management, and Circulation (All Periodicals Publications Except Requester Publications)

1 Publication Title	2. Publication Number	3 Filing Date
The Journal of Philosophy	0 0 2 2 - 3 6 2 X	9/26/2007
4 Issue Frequency	**5 Number of Issues Published Annually**	**6 Annual Subscription Price**
Monthly		Individual $45 00 Institution $100 00

7 Complete Mailing Address of Known Office of Publication *(Not printer) (Street, city, county, state, and ZIP+4®)*
The Journal of Philosophy, Columbia University
1150 Amsterdam Ave, MC 4972
New York, NY 10027
Contact Person: Pam Ward
Telephone (Include area code): 212 666 4419

8. Complete Mailing Address of Headquarters or General Business Office of Publisher *(Not printer)*
Same

9 Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor *(Do not leave blank)*

Publisher (Name and complete mailing address)
The Journal of Philosophy, Inc
1150 Amsterdam Avenue, MC 4972
New York, New York 10027

Editor (Name and complete mailing address)
Arthur Danto, President
The Journal of Philosophy Inc , 1150 Amsterdam Ave, MC 4972
New York, New York 10027

Managing Editor (Name and complete mailing address)
John Smylie, Managing Editor
The Journal of Philosophy Inc , 1150 Amsterdam Ave, MC 4972
New York, New York 10027

10. Owner *(Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a nonprofit organization, give its name and address.)*

Full Name	Complete Mailing Address
The Journal of Philosophy, Inc.	Columbia University
	1150 Amsterdam Avenue
	MC 4972
	New York, New York 10027

11 Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box → ☒ None

Full Name	Complete Mailing Address

12. Tax Status *(For completion by nonprofit organizations authorized to mail at nonprofit rates) (Check one)*
The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes:
☒ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months *(Publisher must submit explanation of change with this statement)*

PS Form 3526, September 2007 *(Page 1 of 3 (Instructions Page 3))* PSN 7530-01-000-9931 PRIVACY NOTICE: See our privacy policy on www.usps.com

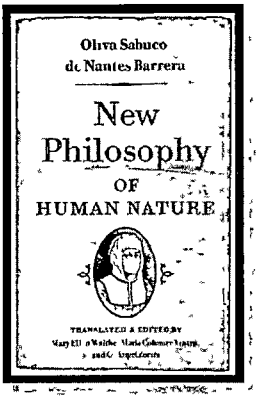
13 Publication Title	14 Issue Date for Circulation Data Below: September 2007	
15 Extent and Nature of Circulation	**Average No. Copies Each Issue During Preceding 12 Months**	**No. Copies of Single Issue Published Nearest to Filing Date**
a. Total Number of Copies *(Net press run)*	3500	3500
b. Paid Circulation *(By Mail and Outside the Mail)* (1) Mailed Outside-County Paid Subscriptions Stated on PS Form 3541 (Include paid distribution above nominal rate, advertiser's proof copies, and exchange copies)	700	674
(2) Mailed In-County Paid Subscriptions Stated on PS Form 3541 *(Include paid distribution above nominal rate, advertiser's proof copies, and exchange copies)*	2400	2247
(3) Paid Distribution Outside the Mails Including Sales Through Dealers and Carriers, Street Vendors, Counter Sales, and Other Paid Distribution Outside USPS®		
(4) Paid Distribution by Other Classes of Mail Through the USPS (e.g. First-Class Mail®)		
c. Total Paid Distribution *(Sum of 15b (1), (2), (3), and (4))*	3100	2921
d. Free or Nominal Rate Distribution *(By Mail and Outside the Mail)* (1) Free or Nominal Rate Outside-County Copies included on PS Form 3541		
(2) Free or Nominal Rate In-County Copies included on PS Form 3541		
(3) Free or Nominal Rate Copies Mailed at Other Classes Through the USPS (e.g. First-Class Mail)		
(4) Free or Nominal Rate Distribution Outside the Mail *(Carriers or other means)*		
e. Total Free or Nominal Rate Distribution *(Sum of 15d (1), (2), (3) and (4))*	150	150
f. Total Distribution *(Sum of 15c and 15e)* ▶	3250	3071
g. Copies not Distributed *(See Instructions to Publishers #4 (page #3))* ▶	250	429
h. Total *(Sum of 15f and g)* ▶	3500	3500
i. Percent Paid *(15c divided by 15f times 100)* ▶	95 45	95 12

16 Publication of Statement of Ownership
☒ If the publication is a general publication, publication of this statement is required. Will be printed in the 9/2007 issue of this publication. ☐ Publication not required

17 Signature and Title of Editor, Publisher, Business Manager, or Owner — Date: 9/30/2007
John Smylie, Managing Editor

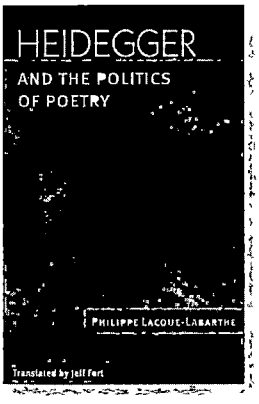
I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including civil penalties).

UNIVERSITY OF ILLINOIS PRESS



New Philosophy of Human Nature
Neither Known to nor Attained by the Great Ancient Philosophers, Which Will Improve Human Life and Health
OLIVA SABUCO de NANTES BARRERA
Translated and Edited by Mary Ellen Waithe, Maria Colomer Vintro, and C Angel Zorita

In 1587 Sabuco developed a new theory of human nature as the foundation for her remarkably modern holistic philosophy of medicine This extensively annotated translation features an introduction demonstrating the work's importance to the history of science, philosophy of medicine, and women's studies
Illus Cloth, $50 00



Heidegger and the Politics of Poetry
PHILIPPE LACOUE-LABARTHE
Translated and with an Introduction by Jeff Fort

This volume collects and translates Philippe Lacoue-Labarthe's studies of Heidegger, written and revised between 1990 and 2002 All deal with Heidegger's relation to politics, specifically through Heidegger's interpretations of the poetry of Holderlin
Cloth, $35 00

New in Paperback
International Nietzsche Studies

Nietzsche's Sister and the Will to Power
A Biography of Elisabeth Forster-Nietzsche
CAROL DIETHE

"A valuable contribution for all interested in Nietzsche studies " —Alexandre Kostka, Universite Cergy-Pontoise and Centre Marc Bloch, Berlin
Illus Paper, $19 95

Distributed for Power Publications at the University of Sydney

Reflections on Japanese Taste
The Structure of Iki
KUKI SHUZO
Translated by John Clark, Edited by Sakuko Matsui and John Clark
Illus Paper, $33 00

Japanese Exchanges in Art 1850s to 1930s with Britain, Continental Europe, and the USA
Papers and Research Materials
JOHN CLARK
Illus Paper, $37 00

www.press.uillinois.edu • 800-621-2736

New from Stanford University Press

Giorgio Agamben
Sovereignty and Life

Edited by MATTHEW CALARCO and STEVEN DECAROLI

"One cannot overstate the importance of Agamben's work or the wide respect it enjoys among scholars in a variety of disciplines in the United States and elsewhere This volume contains valuable essays from a variety of prominent scholars, and is the only such collection available in any language "

—Michael Hardt, Duke University

$21 95 paper $60 00 cloth

Figures of the Thinkable

CORNELIUS CASTORIADIS, Translated by HELEN ARNOLD

In this posthumous collection of writings, Cornelius Castoriadis (1922-1997) pursues his incisive analysis of modern society, the philosophical basis of our ability to change it, and the points of intersection between his many approaches to this theme

Meridian Crossing Aesthetics

$24 95 paper $65 00 cloth

Selected Writings

SARAH KOFMAN Edited by THOMAS ALBRECHT with GEORGIA ALBERT and ELIZABETH G ROTTENBERG

Introduction by JACQUES DERRIDA

This fascinating volume is a comprehensive anthology of significant essays and book excerpts by the postwar French philosopher and theorist Sarah Kofman (1934-1994)

Meridian Crossing Aesthetics

$24 95 paper $65 00 cloth

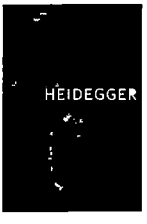
Adorno and Heidegger
Philosophical Questions

Edited by IAIN MACDONALD and KRZYSZTOF ZIAREK

"This fine collection of essays by an important and diverse group of experts brings us to the next stage of evaluating Heidegger and Adorno's work "

—Stephen H Watson, University of Notre Dame

$21 95 paper $60 00 cloth



A Childhood Memory by Piero della Francesca

HUBERT DAMISCH Translated by JOHN GOODMAN

"This book is a brilliant postmodern discussion, grounded in art history, of an enigmatic and perennially fascinating Renaissance painter and his art "

—Ellen Handler Spitz, author of *Art and Psyche* and *Museums of the Mind*

Cultural Memory in the Present

$19 95 paper $55 00 cloth

AVAILABLE IN NOVEMBER 2007

Foucault Beyond Foucault
Power and Its Intensifications since 1984

JEFFREY T NEALON

"*Foucault Beyond Foucault* is the first major renovation of the critical representation of Foucault's system in the past twenty years Jeffrey Nealon successfully challenges the critical prejudices and assumptions that have defined Foucault's legacy, particularly in the North American academy, and stakes out new terrain "

—Gregg Lambert, Syracuse University

$21 95 paper $55 00 cloth

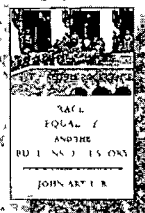
Stanford University Press

800.621 2736 www sup org

CAMBRIDGE

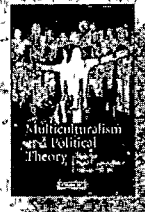
What's New in Philosophy

Race, Equality, and the Burdens of History
John Arthur
$80 00 Hb 978-0 521-87937-8 344 pp
$25 99 Pb 978-0-521-70495 3



Multiculturalism and Political Theory
Edited by Anthony Simon Laden and David Owen
$85 00 Hb 978-0-521-85450 4 432 pp
$32 99 Pb 978 0-521 67090-6

The Theory of Contract Law
New Essays
Edited by Peter Benson
Cambridge Studies in Philosophy and Law
$85 00 Hb 978-0-521-64038-1 364 pp
$39 99 Pb 978 0 521-04132-4



A Theology of Public Life
Charles T Mathewes
Cambridge Studies in Christian Doctrine
$99 00 Hb 978 0-521-83226 7 380 pp

A Philosophy of Political Myth
Chiara Bottici
$80 00 Hb 978-0 521-87655-1 294 pp

Unmodern Men in the Modern World
Radical Islam, Terrorism, and the War on Modernity
Michael J Mazarr
$75 00 Hb 978-0-521-88175 3 312 pp
$19 99 Pb 978 0-521-71291-0

Morality and Political Violence
C A J Coady
$80 00 Hb 978-0-521 56000-9 336 pp
$27 99 Pb 978 0 521-70548-6

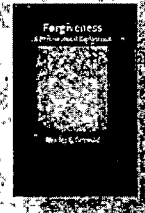
Hellenistic and Early Modern Philosophy
Edited by Jon Miller and Brad Inwood
$37 99 Pb 978-0 521-04244-4 342 pp

Just Health
Meeting Health Needs Fairly
Norman Daniels
$80 00 Hb 978-0-521-87632-2 408 pp
$29 99 Pb 978 0-521-69998-3

Media Argumentation
Dialect, Persuasion and Rhetoric
Douglas Walton
$80 00 Hb 978-0-521-87690-2 396 pp
$28 99 Pb 978 0-521-70030-6

Object and Property
Arda Denkel
Cambridge Studies in Philosophy
$45 00 Pb 978-0 521-04209-3 274 pp

Forgiveness
A Philosophical Exploration
Charles Griswold
$80 00 Hb 978-0-521-87882-1 272 pp
$21 99 Pb 978 0 521-70351-2



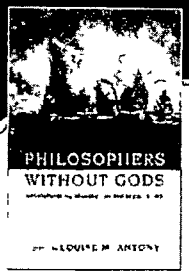
Kant on Beauty and Biology
An Interpretation of the 'Critique of Judgment'
Rachel Zuckert
Modern European Philosophy
$99 00 Hb 978-0-521-86589-0 424 pp

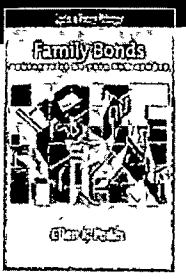
Prices subject to change

www.cambridge.org/us/philosophy

CAMBRIDGE UNIVERSITY PRESS

New from OXFORD





Philosophers without Gods
Meditations on Atheism and the Secular Life

Edited by LOUISE M ANTONY

These highly engaging personal essays capture the marvelous diversity to be found among atheists, providing a portrait that will surprise most readers *Philosophers Without Gods* demonstrates convincingly, with arguments that date back to Plato, that morality is independent of the existence of God Collectively, these essays highlight the richness of atheistic belief—not only as a valid alternative to religion, but as a profoundly fulfilling and moral way of life

2007 336 pp $28 00

Family Bonds
Genealogies of Race and Gender

ELLEN K. FEDER

Ellen Feder's monograph is an attempt to think about the categories of race and gender together Her interdisciplinary work will be of interest to feminist philosophers and theorists because it plays into a recent expansion of interest in the family, as well as to literary scholars of Foucault, to scholars of race and race theory, and to other feminist scholars in political science, sociology, anthropology, and cultural studies

2007 160 pp , 3 halftones, 1 line illus
cloth $74 00 paper $29 95

Self-Transformations
Foucault, Ethics, and Normalized Bodies

CRESSIDA J HEYES

The subject of normalization and its relationship to sex/gender is a major one in feminist theory Heyes' book is unique in her masterful use of Foucault, its clarity, and its sophisticated mix of the theoretical and the anecdotal It will appeal to feminist philosophers and theorists
(Studies in Feminist Philosophy)

2007 176 pp cloth $99 00 paper $29 95

Beyond Reduction
Philosophy of Mind and Post-Reductionist Philosophy of Science

STEVEN HORST

A positive, alternative approach to mind which sees the philosophy of mind, like the sciences, as cognitive enterprises trying to model local features of the world and ourselves
(Philosophy of Mind Series)

2007 240 pp 5 illus $49 95

Medieval Philosophy
A New History of Western Philosophy, *Volume 2*

ANTHONY KENNY

Sir Anthony Kenny continues his magisterial new history of Western philosophy with a fascinating guide through more than a millennium of thought from 400 AD onwards, charting the story of philosophy from the founders of Christian and Islamic thought through to the Renaissance Kenny's lucid and stimulating history will become the definitive work for anyone interested in the people and ideas that shaped the course of Western thought

2007 352 pp 35 b/w illus paper $19 95

OXFORD
UNIVERSITY PRESS

Prices are subject to change and apply only in the US. To order, please call 1-800-451-7556. In Canada, call 1-800-387-8020. Visit our web site at **www.oup.com/us/**

ANNOUNCING:

THE JOURNAL OF PHILOSOPHY

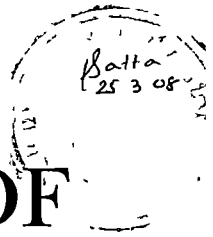
ONLINE

The Journal of Philosophy has added
a print and online subscription option for institutions

SUBSCRIPTION OPTIONS

Libraries and Institutions (print and online)	$180 00
Libraries and Institutions (print only)	$100 00
Individuals	$45 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Please consult our website at www journalofphilosophy org
for more details



THE JOURNAL OF PHILOSOPHY

VOLUME CIV, NUMBER 10
OCTOBER 2007

page		
489	*Conditions for Evolution by Natural Selection*	Peter Godfrey-Smith
517	*An Imperative Theory of Pain*	Colin Klein
	BOOK REVIEWS	
533	Bernard Williams *In the Beginning Was the Deed Realism and Moralism in Political Argument*	Robert B Pippin
540	Bernard Williams *The Sense of the Past Essays in the History of Philosophy*	Jonathan Barnes
546	Bernard Williams *Philosophy as a Humanistic Discipline*	Jonathan Lear

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Wolfgang Mann, Mary Mothersill, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C Varzi *Consulting Editors* David Albert, James T Higginbotham, Robert May, Charles D Parsons, Wilfried Sieg *Managing Editor* John Smylie

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President*, Arthur C Danto, *Vice President*, Akeel Bilgrami, *Secretary*, Michael J Mooney, *Treasurer*, Liaquat Ahamed, *Other Trustees* Lee Bollinger, Leigh S Cauman, Barbara Gimbel, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

You may also visit our website at www journalofphilosophy org

THE JOURNAL OF PHILOSOPHY

2007

SUBSCRIPTIONS (12 issues)

Individuals	$45 00
Libraries and Institutions	$100 00
Libraries and Institutions (print and online)	$180 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for advertising rates, ad space is limited, so ad reservations are required

Published monthly as of January 1977, typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027 (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U S copyright law All rights reserved For more information about our copyright policies, please consult our January issue or contact us

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc , MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027 Periodicals postage paid at New York, NY, and other mailing offices

POSTMASTER. Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027

© Copyright 2007 by the Journal of Philosophy, Inc

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME CIV, NO. 10, OCTOBER 2007

CONDITIONS FOR EVOLUTION BY NATURAL SELECTION*

Both biologists and philosophers often make use of simple verbal formulations of necessary and sufficient conditions for *evolution by natural selection* (ENS) Such summaries go back to Darwin's *Origin of Species* (especially the "Recapitulation"), but recent ones are more compact [1] Perhaps the most commonly cited formulation is due to Richard Lewontin [2] These summaries tend to have three or four conditions, where the core requirement is a combination of *variation, heredity,* and *fitness differences* The summaries are employed in several ways First, they are often used in pedagogical contexts, and in showing the coherence of evolutionary theory in response to attacks from outside biology Second, they are important in discussions of extensions of evolutionary principles to new domains, such as cultural change The summaries also have intrinsic scientific and philosophical interest as attempts to capture some core principles of evolutionary theory in a highly concise way

Despite their prominence, both the proper formulation and status of these summaries are unclear Standard formulations are subject to counterexamples, and their relations to formal models of evolutionary change are not straightforward Here I look closely at these verbal summaries, and at how they relate to formal models Are the summaries merely rough approximations that have no theoretical role of their own? Perhaps they could operate as theoretical state-

*I am grateful to David Haig, Patrick Forber, Samir Okasha, Kim Sterelny, and Michael Weisberg for comments on an earlier draft

[1] Darwin, *On the Origin of Species by Means of Natural Selection* (Facsimile of 1859 edition, Cambridge Harvard, 2001)

[2] See Lewontin, "The Units of Selection," *Annual Review of Ecology and Systematics,* 1 (1970) 1–18

0022-362X/07/0410/489–516 © 2007 The Journal of Philosophy, Inc

ments in Darwin's time, but have now been superseded by more exact treatments

I will look at three families of problem cases, and argue that each motivates different conclusions One set of cases, involving the role of age-structure in populations, is best addressed by regarding the verbal summaries as *idealized* in a particular way A second set of cases, involving heredity, show a role for *approximation* A third set of cases, involving random genetic drift and related phenomena, reveal a way in which a verbal summary, properly formulated, can have a more positive theoretical role These summaries can be used to say things that cannot be said, in suitably general form, via existing formal models At the end of the paper, I offer two new formulations of the traditional three-part summary, guided by a distinction between two roles such formulations can play

I STANDARD FORMULATIONS OF THE CONDITIONS

Perhaps the most commonly cited summary of ENS is due to a 1970 discussion by Lewontin

> As seen by present-day evolutionists, Darwin's scheme embodies three principles
>
> 1 Different individuals in the population have different morphologies, physiologies, and behaviors (phenotypic variation)
> 2 Different phenotypes have different rates of survival and reproduction in different environments (differential fitness)
> 3 There is a correlation between parents and offspring in the contribution of each to future generations (fitness is heritable)
>
> These three principles embody the principle of evolution by natural selection While they hold, a population will undergo evolutionary change (*ibid*, p 1)

Though often cited, this formulation is awkward in several ways, and later formulations by Lewontin and others do not follow it closely In particular, it is not usually seen as necessary for ENS that *fitness* be heritable There are simple cases where fitness is not heritable, but a phenotypic trait is heritable, and that is enough for the trait to evolve Suppose the tall individuals in generation 1 have more offspring than the short ones, and height is heritable Then there will be change from generation 1 to 2, even if there are no fitness differences at all in generation 2 Here fitness is not heritable but height is, and that is enough for height to evolve

This problem with Lewontin's 1970 formulation does become an advantage with respect to one case discussed below But when in this paper I refer to Lewontin's version of the recipe, I will usually mean

the following summary, slightly modified from a 1980 discussion [3] Note that in the summaries quoted below in this section, I have changed the symbols used to number the authors' conditions, for ease of reference below

A sufficient mechanism for evolution by natural selection is contained in three propositions

(L1) There is variation in morphological, physiological, or behavioral traits among members of a species (the principle of variation)
(L2) The variation is in part heritable, so that individuals resemble their relations more than they resemble unrelated individuals and, in particular, offspring resemble their parents (the principle of heredity)
(L3) Different variants leave different numbers of offspring either in immediate or remote generations (the principle of differential fitness)

A more elaborate formulation is offered by John Endler [4]

Natural selection can be defined as a *process* in which

If a population has

(E1) variation among individuals in some attribute or trait *variation,*
(E2) a consistent relationship between that trait and mating ability, fertilizing ability, fertility, fecundity, and, or, survivorship *fitness differences,*
(E3) a consistent relationship, for that trait, between parents and their offspring, which is at least partially independent of common environmental effects *inheritance*

Then

(E4) the trait frequency distribution will differ among age classes or life-history stages, beyond that expected from ontogeny,
(E5) if the population is not at equilibrium, then the trait distribution of all offspring in the population will be predictably different from that of all parents, beyond that expected from conditions E1 and E3 alone

Conditions E1, E2, and E3 are necessary and sufficient for natural selection to occur, and these lead to deductions E4 and E5 As a result of this process, but not necessarily, the trait distribution may change in a predictable way over many generations

[3] See Lewontin, "Adaptation," reprinted in Richard Levins and Lewontin, *The Dialectical Biologist* (Cambridge Harvard, 1985), on p 76 The main modification is that Lewontin said "and" in line 1, but I assume he meant "or"

[4] See Endler, *Natural Selection in the Wild* (Princeton University Press, 1986), p 4

Lastly, here is an example of a textbook presentation, by Mark Ridley[5]

> Natural selection is easiest to understand, in the abstract, as a logical argument, leading from premises to conclusion The argument, in its most general form, requires four conditions
>
> (R1) Reproduction Entities must reproduce to form a new generation
> (R2) Heredity The offspring must tend to resemble their parents roughly speaking, "like must produce like "
> (R3) Variation in the individual characters among the members of the population
> (R4) Variation in the *fitness* of organisms according to the state they have for a heritable character In evolutionary theory, fitness is a technical term, meaning the average number of offspring left by an individual relative to the number of offspring left by an average member of the population This condition means that individuals in the population with some characters must be more likely to reproduce (i e , have higher fitness) than others
>
> If these conditions are met for any property of a species, natural selection automatically results If any conditions are not met, natural selection does not occur

First it is necessary to look at some questions about the intended *role* of these summaries There is an ambiguity in the idea of giving "necessary and sufficient conditions for ENS " The aim may be to describe conditions that will *produce* ENS (where we know what ENS is) Or the aim may be giving conditions for some process *being a case* of ENS So there is a distinction between *constitutive* and *causal* questions to make here (Similarly, giving necessary and sufficient conditions for *becoming* pregnant must be distinguished from giving necessary and sufficient conditions for *being* pregnant)

The usual aim of those offering conditions for ENS seems to be answering both kinds of question The summaries describe a situation in which a certain kind of change will occur, and the entire process is identified with ENS The standard summaries explain what ENS is by giving a *recipe* for ENS But it is also possible to give a summary without giving a recipe We might describe ENS as a temporally extended process of a certain kind, without asserting a tight dependence relation between stages of the process The initial stages of the process may not invariably (or even reliably) be followed by the latter stages, but if the initial stages *are* followed by the latter stages, then we have a case of ENS

[5] See Ridley, *Evolution* (Malden, MA Blackwell, 1996, 2nd ed), pp 71–72

Summaries of ENS, ever since Darwin, have often been presented in the form of recipes for change It is often seen as a strength of evolutionary theory that its core mechanism (or one of its core mechanisms) has a straightforward predictive character More specifically, I take it that the usual aim is to give conditions that are *sufficient ceteris paribus* for a certain kind of change occurring (The status of the *ceteris paribus* clause will be discussed later) Once this is determined to be the aim, there is a further ambiguity, however The recipes are usually interpreted as saying that whenever we have variation, heritability, and fitness differences with respect to some trait in a population, change ensues But some of the recipes, including Lewontin's formulations, might be instead read as saying that whenever a population features a *general* tendency to exhibit variation, heritability, and fitness differences, then *some* traits will change Here I will interpret the recipes in the former, trait-specific way

I will finish this section with brief comments about topics that are not discussed below First, a few summaries require that variation be random [6] I assume that no randomness condition should be included Darwinian evolution can occur on variation that is directional, even adaptively "directed " In these cases natural selection may have less explanatory importance than it has when variation is random, but it can still exist Second, there is no distinction made in summaries above between *natural* and *artificial* selection This, also, will not be treated as a problem, as a dichotomy between natural and artificial selection is not of theoretical importance within biology itself [7] Third, some discussions of the core of Darwinism focus on *cumulative* selection, but I avoid any such restriction and focus on a more inclusive category [8] Lastly, there is an alternative foundational description of ENS that uses the idea of a "replicator " But when intended as a fully general description that can function as an alternative to the standard summaries, this approach fails [9] Though those who advocate this approach usually do not define replicators carefully, this concept involves some absolute (though vague) notion of faithful transfer of structure across generations But what is needed for ENS is parent/

[6] See, for example, Joseph Fracchia and Lewontin, "Does Culture Evolve?" *History and Theory*, XXXVIII (1999) 52–78

[7] Many cases of "natural" selection, whether humans are involved or not, feature a key role for preference and choice as a causal factor

[8] See Kim Sterelny and Paul Griffiths, *Sex and Death An Introduction to Philosophy of Biology* (Chicago University Press, 1999), chapter 2

[9] For expositions of this approach, see Richard Dawkins, *The Selfish Gene* (New York Oxford, 1976), and David Hull, "Individuality and Selection," *Annual Review of Ecology and Systematics*, XI (1980) 311–32 For the criticisms summarized in this paragraph, see my "The Replicator in Retrospect," *Biology and Philosophy*, XV (2000) 403–23

offspring similarity understood in a *comparative* way (This point will survive the problems and modifications discussed below.) It is sufficient for ENS (given other conditions) that parent and offspring be *more* similar than randomly chosen individuals of different generations. So any absolute degree of parent/offspring similarity (except 100%) will be sufficient in some contexts and insufficient in others. Replicators are also apparently meant to be asexual in their "transfer of structure," and ENS should surely be possible with sex at the focal level—without sexual reproduction being reducible to asexual replication of genes, for example. As a consequence, replicators are not necessary for ENS.

II. BIRTHS AND DEATHS

I now turn to cases that cause problems for the standard summaries. The first is not a puzzle case *per se*, but a routine phenomenon whose proper categorization is unclear, and that introduces more important cases.

> *Case 1, Culling.* We have a population of individuals, of types *A* and *B*. Across a time interval, some individuals die while some remain alive. The frequencies of types change over the interval, as *B* individuals die at a higher rate than *A*. The higher death rate in *B* is due to their inferior ability to fight off disease.

No one reproduces, so the Lewontin 1980 and Ridley summaries preclude this from being a case of ENS. Endler's formulation and Lewontin's 1970 account allow that survivorship differences alone count as fitness differences, but if culling alone counts for ENS then heritability is not strictly necessary.

A natural response is to deny that this is a case of ENS. Pure culling is *part* of a process of ENS, but not sufficient alone. This response is largely right, but it is worth going through some arguments that can be made on each side.

Those who think culling should count as ENS might note that "change in gene frequencies" is supposed to be sufficient for evolution in general. That is a textbook criterion (due originally to Theodosius Dobzhansky). If so, change in gene frequencies due to culling on the basis of phenotype should be enough for ENS. In reply, it may be argued that this shows a misunderstanding of the idea of "change in gene frequencies." The standard criterion for evolution is supposed to be understood as change in gene frequencies *across generations*. So to see whether or not there is ENS, we must wait until the individuals in Case 1 reproduce.

But this reply leads to trouble—to the unraveling, in effect, of an idealization that lies in the background of many discussions. We are

supposed to wait until *who* reproduces? All of the population? Only some of them? All of those who will reproduce eventually?

To air these issues more fully, I introduce Case 2

> *Case 2, Different Generation Times* We have a population of individuals, of types *A* and *B* Every individual alive at the beginning of some time interval fissions to produce two offspring of the same type as the parent Later, all individuals do the same thing again, and again But *A* individuals cycle through this process faster than *B* individuals do, owing to their more efficient metabolism So more *A*'s are produced, and at the end of the time interval the frequencies of the types have changed

This seems to be clearly a case of ENS There is a change in the frequencies of types, due to variation in reproductive capacities and faithful inheritance of type There are no differences in the population with respect to the *number* of offspring produced by different individuals, however So there are no differences between the two types in the number of offspring produced per individual of that type The only differences between the types concern the rate at which new individuals are produced per unit of *time*[10]

The Lewontin 1980 and Ridley summaries clearly exclude this case (Endler will be discussed below) Lewontin and Ridley explicitly treat fitness as the *number* of offspring produced by an individual (or the number produced on average by individuals of a given type)[11] All the individuals in Case 2 have two offspring—eventually—and the "eventually" versus "quickly" distinction has no place in these summaries

This is not to say that Case 2 is a problem for evolutionary theory itself This is a simple case of an "age-structured population," and there are detailed models of such cases[12] I will sketch the simplest kind of analysis that would be given

The crucial difference is that we now think of reproduction as occurring in time For simplicity I will treat time in a discrete way,

[10] The reference to "remote generations" in the Lewontin formula does not help All individuals have the same number of grand-offspring and great-grand-offspring as each other, though some individuals take longer than others to achieve them

[11] Lewontin's 1970 recipe uses the term "rate" which is ambiguous with respect to units, but certainly may include measures per unit of time

[12] For helpful summaries, see James Crow, *Basic Concepts in Population, Quantitative, and Evolutionary Genetics* (New York Freeman, 1986), chapter 6, and Jonathan Roughgarden, *Theory of Population Genetics and Evolutionary Ecology An Introduction* (Upper Saddle River, NJ Prentice Hall, 1979), chapter 18 For a detailed account, see Brian Charlesworth, *Evolution in Age-Structured Populations* (New York Cambridge, 2004)

measured in days The two types (A and B) each have an "$l(x)$ schedule," which specifies how likely an individual of that type is to reach age x, and an "$m(x)$ schedule," which specifies how many offspring an individual of that type will have at age x

We first think of the As and Bs as forming two subpopulations Each subpopulation will reach a stable distribution of ages and then grow multiplicatively, by a factor of λ_A and λ_B per day respectively To determine these rates of increase for the two types, we solve the following equation separately for each type

$$1 = \sum_x \lambda^{-x} l(x) m(x) \tag{1}$$

Let us assume the following $l(x)$ and $m(x)$ schedules, which satisfy the description given for Case 2 above,

Type A $l(1) = 1$, $l(2) = 0$, $l(3) = 0$

$m(1) = 2$, $m(2) = 0$, $m(3) = 0$

Type B $l(1) = 1$, $l(2) = 1$, $l(3) = 0$

$m(1) = 0$, $m(2) = 2$, $m(3) = 0$

We find that $\lambda_A = 2$ and $\lambda_B = \sqrt{2}$ These numbers can be used to get a representation of the rate at which the frequency of A will grow relative to B, in the total population

So it is possible in Case 2 to assign to each type a fitness-*like* number, that will predict what will happen in the population These are the *rates of increase* of the two types, measured with respect to time I return to the status of these parameters below, but the immediate point is that we see that several recipes for ENS are making a tacit *idealization* They treat all cases of ENS as if they occurred in populations in which generations are *nonoverlapping* and *synchronized* across the entire population (This is often called a "discrete generation" model, though this should not be confused with treating time in a discrete way, as I did above) The same idealization has also been operating, often unacknowledged, in much of the literature on the "propensity view of fitness "[13]

The term "idealization" is a controversial one in philosophy of science I understand idealization as involving the *imagined modification* of a real system, usually in the direction of simplicity An idealized description is one that is straightforwardly true of a fictional relative

[13] See Susan Mills and John Beatty, "The Propensity Interpretation of Fitness," *Philosophy of Science*, XLVI (1979) 263–88

of the real system, and may also be taken (in many cases) to be approximately true of the more complex real system [14] The present example involves a special kind of idealization, however Some organisms *do* have nonoverlapping generations synchronized across the population—annual plants such as basil do, many insects, and some others But most organisms do not In a case like the human population, the notion of a "generation" has no meaning as a population-level, as opposed to individual-level, phenomenon So summaries of ENS given in the style of Lewontin and Ridley, which treat fitness as the number of offspring produced (or the expected number) can be applied *literally* to annual plants, many insects, and some other organisms But to most organisms they must be applied in an indirect way, via an idealization A summary that treats fitness in terms of numbers, ignoring the role of time, can only be seen as describing an *imagined simpler relative* of the processes of ENS in those cases

At this point, it might be thought that the right response is to modify the summaries so that they use rates of increase And the simplest "discrete generation" cases can indeed be treated as special cases of an evolutionary process with age-structure However, models that predict change with rates of increase (like λ_A and λ_B above) make their own idealizations Above we assumed asexual reproduction, and $l(x)$ and $m(x)$ schedules that stay fixed as the population grows Once the population is sexual, and creates new individuals by combining contributions from two parents, we cannot represent the rate at which a type increases in terms of its *own* survival and reproduction schedules This is because any type also produces other types, and is produced by them—if, indeed, discrete "types" exist at all in the population, which may instead contain individuals who vary only quantitatively We also assumed that the population was in a stable age distribution, though most populations will be knocked out of this distribution by natural selection itself, along with other factors [15]

So those who want an exact formal treatment face a choice between idealizations The consensus among modelers seems to be that in many empirical cases, a good approximate description can be achieved by assuming either a discrete generation model or an age-structured

[14] See Martin Thompson-Jones, "Idealization and Abstraction A Framework," in Thompson-Jones and Nancy Cartwright, eds , *Idealization XII Correcting the Model* (Amsterdam Rodopi, 2005), pp 173–217, and my "Abstractions, Idealizations, and Evolutionary Biology," to appear in A Barberousse, M Morange, and T Pradeu, eds , *Mapping the Future of Biology Evolving Concepts and Theories* (forthcoming)

[15] See Andre Ariew and Lewontin, "The Confusions of Fitness," *British Journal for the Philosophy of Science,* LV (2004) 347–63

model with rates of increase, and these two models converge a lot of the time, especially when selection is weak [16] Thus a modeler can pick and choose between frameworks, according to the case and the purposes at hand But if our aim is formulating necessary and sufficient conditions for ENS itself, which we aim to see as a single kind of real process, then the role of these idealizations is problematic In particular, it becomes impossible to treat ENS as a process that is always "driven" by something like the familiar fitness differences invoked by Lewontin and Ridley [17]

So far, we have seen that there are two ways to approach the formulation of a summary of ENS One way is to make idealizations, and give a summary that will apply literally to some cases and via idealization to others Then it is possible (modulo some complications discussed below) to keep the summary simple, while also specifying a process with the kind of internal causal reliability or predictive power that was discussed in the previous section That is, it becomes possible to give a summary in the form of a recipe

The other approach is to avoid idealization, and try to capture every case In retrospect, we can see that this is what Endler's formulation was doing Endler's summary is expressed *as if* it is describing a recipe with the kind of predictive features seen in Lewontin's, but the formulation is so full of qualifications that it has little predictive power In his "E2" clause Endler lists a number of properties that are *related* to the notion of fitness—survival, fecundity, fertility, mating ability—but he does not collapse these into a single measure that is taken to be predictive of change He does not say that the "bottom line" for ENS is differences in expected number of offspring, or differences in rate of increase, or something else If there is no "bottom line," Endler is leaving it open that the "mating ability" differences might balance out the "survival" differences, for example, to yield no evolutionary change

So the Lewontin/Ridley approach is to idealize, while the Endler approach is to avoid idealization at the expense of predictive "punch " And what is then surprising is that none of Ridley, Lewontin, or Endler *say* that this is what they are doing Ridley and Lewontin do not confess to idealization, and Endler hangs onto the language of definite pre-

[16] See Crow, *op cit*, p 174

[17] There are other measures besides the ones discussed here But Charlesworth's authoritative survey concludes that in an age-structured sexual population "no single parameter can be regarded as the fitness of a genotype with arbitrary selection intensities" (*op cit*, p 136)

diction even when the details of his formulation let much of the predictive air out

I will make a few more comments on the relation between the idealizing and nonidealizing approaches, and then move to a new set of problem cases First, another role that the idealization to discrete generations achieves is the establishment of a minimal unit of evolutionary change, which becomes change across a single generation Then it is easy to say that Case 1, pure culling, does not count as ENS But once we are thinking of an age-structured population, there is no nonarbitrary minimal unit Then we just have shorter and longer intervals of time (The Price equation approach to modeling evolution, which will be discussed below, embodies such a picture) An interval too short will be one in which nothing of interest can happen, and an interval too long is one that might be outside the domain of micro-evolutionary theory (the theory of change within a population) altogether But within those boundaries, there is a lot of freedom That is not itself a problem Modelers like to think in terms of minimal units of change, but that is purely for convenience There are ways to avoid wrangling about the status of Case 1 within such a picture One can say that the *paradigm* cases of ENS include reproduction and extend over many typical generation lengths for the organisms in question If we ask what is the *minimal fragment* of such a process that counts as ENS, the question can be dismissed as empty As we get further and further from the paradigm cases, we get further and further from having a fragment that deserves the name "a case of ENS "

We have seen in this section that the standard summaries often engage in idealization, and that avoiding idealization trades off against predictive power Further, the idealizations are present even when the language used by an author does not acknowledge idealization, but seems intended to establish as literal and direct a mode of description as possible

III HEREDITY AND HERITABILITY

All summaries of ENS include a requirement involving the inheritance of traits Lewontin's summaries require that variation be "heritable," where this is understood in terms of a statistical similarity between parent and offspring Endler and Ridley are less specific (see clauses E3 and R2) Here I use the term "heritability" to refer to a family of statistical measures, and "heredity" and "inheritance" to refer in a vaguer way to all phenomena involving parent/offspring similarity and the transmission of traits across generations

The problems encountered in this section are in some ways reminiscent of those in the previous one It is appealing to summarize ENS

by giving a recipe for change, especially if the recipe is a compact one with clearly separable ingredients But summaries that satisfy this goal are not able to handle all cases Other problems we will encounter have no analogues in the previous section

The role of heredity in ENS is often described using metaphors If a population has variation and fitness differences, there will be no change if the population is not disposed to "respond" to selection, and that requires that parental characters are "transmitted" across generations Lewontin's summaries use the concept of heritability to capture this extra ingredient precisely Summaries of this kind also shadow a formal representation of evolution known as the "breeder's equation," which treats change in the mean value of some characteristic in a population as the product of a term representing variation in fitness and a term representing heritability [18]

It is common to distinguish several different senses of heritability, which arise within two frameworks or approaches [19] One approach, which I will call the "fraction of variance" approach, is based on a causal model of inheritance that assumes the presence of genes or something similar to them Heritability is measured as the genetic variance divided by the total phenotypic variance in the population The other, the "regression approach" is more abstract, independent of any causal model of inheritance, and aimed simply at the representation of predictability relations between parents and offspring This second approach will be used here, as it is important not to assume that standard genetic mechanisms of inheritance are present in all cases Heritability will be initially understood as the slope of the linear regression of offspring character on parental character [20] This approach to the "third ingredient" nicely meets the goals discussed above Heritability in this sense is a single number, representing the tendency of population to "respond" to selection Further, it is usually understood and measured independently of fitness differences, so it operates as a distinct ingredient in the recipe, not something conceptually entangled with selection itself

[18] See John Heywood, "An Exact Form of the Breeder's Equation for the Evolution of a Quantitative Trait under Natural Selection," *Evolution*, LIX (2005) 2287–98

[19] See A Jacquard, "Heritability One Word, Three Concepts," *Biometrics*, XXXIX (1983) 465–77, and S Downes, "Heredity and Heritability," *Stanford Encyclopedia of Philosophy* (http //plato stanford edu/entries/heredity/)

[20] Within the fraction of variance approach, both "narrow" and "broad" sense heritabilities can be distinguished When both frameworks (fraction of variance, and regression) are applicable, the sense of heritability measured as the regression of offspring on parent corresponds to the narrow sense

I now turn to some problem cases The symbolism used is as follows In the population of parents, individual i has phenotypic value Z_i This is a quantitatively varying character, such as height The average phenotypic value that the offspring of individual i have, or would have if there were any, is Z'_i In most of this section I assume asexual reproduction, so each individual has just one parent Heritability is measured as the slope of the line that gives the best prediction of Z' values from Z values [21] Absolute fitness is symbolized with W

The first case is introduced in an unpublished manuscript by Robert Brandon [22]

> *Case 3, Biased Inheritance* A population varies with respect to Z (Brandon's example is wealth) Z is heritable, and positively associated with fitness But there is also a tendency for offspring to have a lower Z value than their parents A numerical example is given in Table 1 [23] The population's composition is unchanged across generations, despite variation, fitness differences, and a heritability of 0 74

Table 1 Numerical example of biased inheritance, after Brandon

Individual	Z	W	Z'
1	2	2	2
2	2	2	2
3	2	1	1 67
4	2	1	1 67
5	1 67	1	1 67
6	1 67	1	1 67
7	1 67	1	1 33
8	1 67	1	1 33
9	1 33	1	1 33
10	1 33	1	1 33
11	1 33	0	1 33
12	1 33	0	1 33

Brandon argues that this shows that Lewontin-style summaries assume the absence of *bias* in the inheritance system Although it was said above that heritability is identified with the slope of a regression line predicting Z' from Z, a regression analysis gives us two parameters, the slope of the line and the intercept with the vertical axis Here the intercept is negative, and the influence of this factor exactly

[21] The regression line is the predictor of Z' from Z that minimizes the squared distances of the data points from the line, and is calculated as $Cov(Z', Z)/Var(Z)$ As will become conspicuous, the remainder of this section will use a discrete-generations assumption of the kind discussed in the previous section

[22] Brandon, "Inheritance Biases and the Insufficiency of Darwin's Three Conditions" (manuscript)

[23] Though Z' is generally a mean in this paper, this table should be read so that the top two $Z = 2$ individuals have only offspring whose Z value is 2

counteracts the evolutionary change that would have been predicted from the fitness differences and the heritability alone So if heritability is understood as a regression slope, then at least one extra parameter needs to be taken into account when predicting change

I now introduce a second problem case involving heritability

Case 4, Heritability Fails in the Fit An asexual population contains variation in height There is a positive covariance between height and fitness There is a positive covariance between parental height and offspring height But there is no change in height across generations This is because although taller individuals have more offspring on average, and taller individuals have taller offspring on average, the taller individuals with the high fitness are not the *same* tall individuals as those that have taller offspring The high-fitness tall individuals are not the tall-offspring tall individuals The mean value of Z is unchanged across generations, though there are fitness differences and Z is highly heritable A numerical example is given in Table 2

Table 2 Numerical Example for Case 4

Individual	Z	W	Z'
1	1	1	1
2	1	1	1
3	1	1	1
4	2	1	2
5	2	1	2
6	2	7	1 571
Mean	1 5	2	1 5 (weighted by size of offspring classes)

Once again, the three-part recipe is not sufficient for change (If individual 6 produces 4 offspring of $Z=2$ and 3 of $Z=1$, we get the same population statistics back with a larger population size) The key point is obvious in retrospect If we calculate heritability from the entire parental population, the heritability can be affected by individuals who make little or no contribution to the next generation So there *has* to be some error in a prediction that is made using heritability in this way, unless the pattern of inheritance is the same across the whole population

This, like Brandon's case, is a toy example But both cases illustrate phenomena that are real possibilities One is a directional tendency in the mechanisms producing departure of offspring from parental phenotype The other is a mixed underlying basis for the inheritance of a single trait, so that the fittest individuals do not have the same inheritance patterns as those seen for other values of the trait In each case, a fine-grained analysis removes any appearance of paradox, but that does not invalidate the fact that the population-level criteria used for predicting change have problems

The obvious response to Case 4 is to understand heritability in a way that takes into account fitness differences in the parental generation We do not need to modify the formulation of the recipe itself, but just the interpretation of a key component This response is reasonable [24] This has consequences for how a recipe for evolutionary change is understood, however The original aim was to think of heritability as dispositional property of the parent population that exists independently of the pattern of fitness differences Whether or not heritability should be called a "cause" of anything, the aim was to treat it as a distinct ingredient in a breakdown of explanatory factors If heritability properties are treated as dependent on fitness, we have logically "entangled" two ingredients in the breakdown of factors [25]

Yet another set of problem cases involve the interaction of heredity with *stabilizing selection,* selection that acts to maintain an intermediate value of some trait in the population A simple case is as follows

Case 5 Stabilizing Selection in an Asexual Population (see Figure 1) The population contains short (Z=1), intermediate (Z=2), and tall (Z=3)

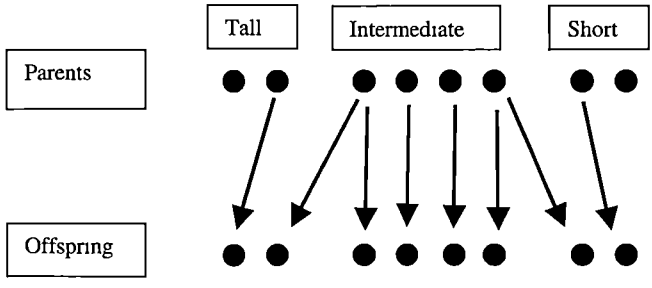


Figure 1 Stabilizing selection in an asexual population [26]

[24] I modified Brandon's numerical example to reduce, but not eliminate, the role of these factors Individuals 11 and 12 in the chart do not reproduce at all, but they are associated dispositionally with a Z' of 1 33 This makes sense if heritability is seen as a dispositional property of the parents independent of fitness, but, of course, the basis of such dispositions may be controversial And we see that there are various ways of thinking of heritability here We could only count the parents that do reproduce, but ignore how much they reproduce Or we could weight each Z_i and Z'_i pair by the fitness of i See Heywood (*op cit*) for an endorsement of fitness-weighted heritabilities

[25] This entanglement is already present in the fact that variation must exist in the population if there are fitness differences and/or heritability The ingredients that it would be attractive to keep separate are heritability and fitness, though This goal exerts real influence, for example, on Okasha's treatment of the problem, as discussed below

[26] If you are like me, you will see a slight optical illusion here, where the lower "flanking" individuals seem pulled in towards the central group Note that there is not supposed to be phenotypic variation within the three classes here

individuals Intermediate individuals are fitter than either extreme Heritability as measured by parent-offspring regression is one, but the parental and offspring generations are phenotypically identical Stabilizing selection is exactly compensated by a dispersing tendency in inheritance seen in the phenotypically intermediate individuals

This case mirrors a category of cases involving sexual reproduction, featuring heterozygote superiority with respect to fitness but not with respect to phenotype That is, an intermediate phenotype is favored by selection and is produced by a heterozygote at one locus (genotype *Aa*), resulting in a stable equilibrium of gene frequencies There is a tendency for short individuals to produce short individuals and tall to produce tall, hence heritability of phenotype, even when the population is in the equilibrium state And there are fitness differences between individuals in this equilibrium state This I will call *Case 6* A graphic representation of an extreme example is given in Figure 2

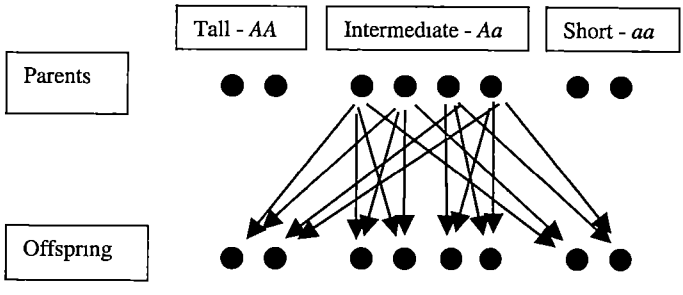


Figure 2 A superior intermediate phenotype produced by a heterozygote (Case 6)

The message of these cases is that heritability is quite a blunt instrument, even when our sole aim is to say whether change will occur or not These cases also highlight a fact about stabilizing selection In talking about selection, it is often said that fitness differences in the population must be "systematic" Stabilizing selection involves systematic fitness differences in one sense, but not in another sense These are cases where we can say in independent terms which phenotypes are the fit ones, and may be able to give ecological reasons why they are fit A golden mean principle may be applicable, for example But in another sense, the fitness differences in these cases are not "systematic" because there is no overall tendency for higher values of Z to be fitter, or less fit, than low values This will be discussed again in the next section

This case also casts some light on the recipes quoted in section II First, Lewontin's 1970 recipe required that *fitness* be heritable, not the

phenotypic trait evolving In some simple cases this seems to give the wrong answer, and Lewontin did not use this condition in his 1980 and later formulations But in Case 6, fitness is not heritable even though phenotype is So Lewontin's 1970 recipe has no problem in that case, and he notes these phenomena in his 1970 discussion However, both fitness and phenotype are heritable in Case 5 [27] Second, Endler's qualifications in his clause E5 are probably intended to handle these sorts of cases

So the use of heritability in specifications of conditions sufficient for change must be seen as involving an approximation Positive heritability is compatible with inheritance biases that can cancel change If heritability is measured without regard for fitness, it is affected by misleading information And the pattern of heredity can interact with stabilizing selection in such a way that despite high heritability and fitness differences, the same population is restored across generations

In some (though not all) of the cases discussed in this section, the problem arises because we stipulate a pattern of heredity that is liable to produce change *on its own,* and selection exactly cancels that tendency, yielding no net change Such cases might reasonably be regarded as special If so, Case 6, which does not have this feature, becomes an important one

In this section I have used the term "approximation" to describe the role of heritability In the previous section I discussed *idealization* How are these phenomena related?

When idealizing in the manner discussed in the previous section, we imagine a *structural modification* to the systems we are interested in, in order to make them easier to describe In these cases involving heritability, that is not what is going on (or at least, not with respect to the features emphasized in this section) Rather, we ignore some possibilities, and are also content with an analysis that makes predictions that are *largely* accurate the rest of the time We are not imagining structural modifications, but merely allowing our descriptions to have a "loose fit" to real-world phenomena We *could* describe the situation as one in which we are "idealizing away from" certain possibilities, but the specific type of imaginative act that was uncovered in section II is not found here

Though an idealized description is only straightforwardly true of an imagined relative of a real system, it will often be approximately true of the more complex real-world system So idealization can yield approximate truth But not all approximately true descriptions involve idealization If I say someone is 6 feet tall when they are only

[27] Here I assume that the same pattern of reproduction continues in the new generation

5 feet 11 inches, this will count as approximately true in many contexts, but I am not idealizing in the sense discussed above

In sum the problems involving age-structure in section II show the role of idealization in standard summaries of ENS, while the problems involving heritability show a role for approximation in a more general sense that need not be understood in terms of idealization

The cases in this section also raise the possibility of appealing to *ceteris paribus* clauses in making sense of the recipes Certainly we must assume that something like a *ceteris paribus* clause is in the background in all claims about the conditions sufficient for evolutionary change to occur A variety of strange events outside the usual domain of evolutionary biology could intervene to prevent change Though I accept that minimal role for a *ceteris paribus* assumption, I doubt that the specific problems raised here are best handled by leaning on this idea We are not confronting problems that involve breakdowns of normal conditions or processes Rather, they are ordinary biological complexities that are conspiring to make it hard to say what we want to say in a compact way

IV HERITABILITY AND THE PRICE EQUATION

In this section I discuss the relation between the traditional summaries and the "Price equation" framework for the abstract representation of evolutionary change Samir Okasha, in an important recent discussion, has claimed that, in the light of Price, the status of the Lewontin conditions must be revised [28] My discussion here draws extensively on Okasha's treatment

One form of the Price equation for representing change is as follows

$$\Delta\bar{Z} = Cov(w,Z) + E_w(\Delta Z) \tag{2}$$

As above, Z is a quantitative character (such as height), and $\bar{Z}$ is its mean in the parental generation The term $\Delta\ \bar{Z}$ is defined as $\bar{Z}_o - \bar{Z}$, where $\bar{Z}_o$ is the mean in the next generation Fitness, symbolized by w, is now a form of relative fitness, in which absolute fitness is divided by mean fitness in the population So each individual i in the parental generation is characterized by its Z_i and w_i, its phenotype and its fitness, and also by Z'_i, the average Z value of its offspring An individual is also characterized by its ΔZ_i, which is its value of $Z'_i - Z_i$ Then $Cov(w, Z)$ is the covariance in the population between Z and fitness $E_w(\Delta Z)$ is the fitness-weighted average of the ΔZ values [29]

[28] Okasha, *Evolution and the Levels of Selection* (New York Oxford, 2005), chapter 1

[29] Derivations of the equation are given in Steven A Frank, *Foundations of Social Evolution* (Princeton University Press, 1998), and Okasha, *op cit*

$Cov(w, Z)$ is zero in the cases of stabilizing selection (Cases 5 and 6) discussed in the previous section. Although there were fitness differences in those cases, there was no overall tendency for either high or low values of Z to have high fitness

Initially, it seems that there is also a simple link between this equation and a Lewontin-style recipe Evolutionary change has been broken down into a term that represents the role of fitness differences, $Cov(w, Z)$, and a term that represents the net role of inheritance, expressed as a correction for "transmission bias" The link can be made more explicit by unpacking the covariance term into the regression of fitness on character ($b_{w,Z}$), and the variance of Z

$$\Delta\bar{Z} = b_{w,Z}Var(Z) + E_w(\Delta Z) \tag{3}$$

So we have terms representing variation, fitness differences associated with phenotype, and the role of inheritance But the term $E_w(\Delta Z)$ is far from a standard heritability measure To locate the role of heritability in the usual sense in equation (2), we must break it down in a more complicated way Combining several equations from Okasha's treatment

$$\Delta\bar{Z} = hCov(w,Z) + Cov(w,e) + a + \bar{Z}(h - 1) \tag{4}$$

Here, h is the regression slope of Z' on Z, without fitness weighting The intercept of that regression is a $Cov(w, e)$ is the covariance between the fitness of each individual i and the "residual" or error for that individual (e_i) when using the regression line $Z'_i = hZ_i + a$ to predict its value of Z'_i So we see via Price that heritability in this familiar sense is only one of *three* factors that matter, concerning inheritance Case (3) above works via the role of a, and Case 4 works because of the role of $Cov(w, e)$ [30]

Okasha argues, more strongly, that in the light of the Price equation, the Lewontin conditions are structurally problematic His argument is as follows It is desirable to re-write the Price equation in a way that removes the role of fitness differences (reflected in the weighting of the average) from the second term on the right hand side, the one representing the role of the inheritance system That yields this formula [31]

$$\Delta\bar{Z} = Cov(w,Z') + E(\Delta Z) \tag{5}$$

[30] And also a nonzero intercept a

[31] This formula is used in Sean Rice, *Evolutionary Theory Mathematical and Conceptual Foundations* (Sunderland Sinauer, 2004), and Heywood (*op cit*)

But once we have this version of the equation, we see that "what is *really* required for there to be evolution by natural selection is for $Cov(w, Z')$ to be non-zero, that is, for an entity's fitness to correlate with the average character of its *offspring* This is the fundamental condition " (*op cit*, p 37, some symbols changed) That is what tells us when there will be change over and above that produced by the inheritance system alone The Lewontin recipe, in this analysis, tries to capture this "fundamental condition" with two others, that $Cov(w, Z)$ and $Cov(Z, Z')$ both be nonzero But $Cov(w, Z')$ cannot be determined from these In effect, the Lewontin conditions treat covariance as a transitive relation, which it is not

Though this analysis is very illuminating, I do not accept the conclusions Okasha draws I first introduce a problem case If "what is *really* required for there to be evolution by natural selection is for $Cov(w, Z')$ to be nonzero" then this is a case of evolution by natural selection

> *Case 7, $Cov(w, Z')$ Positive with No Variation in Z* Some individuals in the parental generation have more offspring than others, and the offspring of these individuals have higher values of Z than the mean value in the parental population The mean value of Z is then higher in the offspring generation But the reproductively successful individuals did not *themselves* have high values of Z In fact, everyone in the parental generation had the same value of Z

This certainly does not look like a case of ENS, at least with respect to Z However, it is also underspecified Why did the high Z' individuals have more offspring, if there was no variation in Z? Given the story told, there seem to be two options One is pure accident, and the other is hitchhiking, the individuals with high values of Z' were favored by selection on some other trait Both of these cases must be distinguished from natural selection acting on Z itself So far, the Lewontin recipes (although not the Endler or Ridley recipes) have said nothing to help us distinguish selection from random drift or hitchhiking But as I will argue in the next section, the way to make sense of these distinctions is *not* to take Z, the parental phenotype, out of the picture, as $Cov(w, Z')$ does

Second, Okasha's claim that $Cov(w, Z')$ is the fundamental condition is based on the claim that equation (5) has a kind of primacy in its representation of the factors responsible for change, in particular a superiority over equation (2) This can be questioned on independent grounds Okasha's aim is to treat the second term on the right hand side of the equation, which handles the role of "transmission bias," in a way independent of the effect of fitness differences That is apparently a desirable separation However, the removal of fitness

differences from this term introduces an effect of the *inheritance* system on the representation of the role of *fitness* differences, in the new first term on the right hand side Consequently, I do not agree that equation (5) gives us a true separation of the role of fitness and inheritance

Another way to look at it is to compare yet another Pricean breakdown of change

$$\Delta\bar{Z} = Cov(w,Z) + Cov(w,\Delta Z) + E(\Delta Z) \quad (6)$$

The first term on the right hand side represents fitness differences only, the far right term concerns the inheritance system only, and the middle term combines both That middle term can either be assimilated into the $Cov(w, Z)$ term, yielding equation (5), or assimilated into the $E(\Delta Z)$ term, yielding equation (2) Either way, we get a mixture of the role of inheritance and fitness somewhere So I do not think either (2) or (5) is superior to the other, in a principled sense [32]

In sum, I accept that the Pricean analysis is very informative about the status and workings of the traditional recipes, as Okasha claims It helps us understand the respects in which those recipes are approximations However, I am not convinced by Okasha's arguments about the need for a *restructuring* of summaries of ENS

I close the section with two more general remarks on heredity First, throughout this section I have assumed that we are dealing with a trait that varies quantitatively This is needed for the measurement of variances and covariances But other traits are more naturally represented as occurring in discrete types without well-defined distances between them These cases can be handled within the quantitative framework via some shoe-horning, but there are probably better ways to represent them [33] Once again, it is easy to find ourselves analyzing a particular subset of the cases and then treating the outcome of the analysis as applying straightforwardly to all

[32] Okasha also gives an argument based on counterfactuals, but I think it has similar problems He argues that standard ways of assessing counterfactuals imply that $E(\Delta Z)$ expresses the change in Z that would have occurred if selection was wholly absent, and this shows that $Cov(w, Z')$ measures "the difference made" by selection But if we consider the other counterfactual, imagining inheritance bias wholly absent, we do *not* find that $Cov(w, Z')$ measures the change that would then result So again, I do not think that equation (5) gives us a superior breakdown

Suppose that only one individual in the parental generation exhibits a tendency to biased transmission But that individual has zero fitness Then we can say *either* that this was a case where transmission bias played no role (talking the language of equation 2), *or* that this was a case where transmission bias was present in the system but was counteracted by $Cov(w, Z')$

[33] This is in effect what mutation parameters do

Second, we should note one other fact about the treatment of heritability in terms of regression A regression can have any value between plus and minus infinity On the "fraction of variance" approach, heritabilities are between zero and one (inclusive) If we assume standard genetic mechanisms, heritabilities measured by regressions lie between zero and one and the two frameworks coincide But if our aim is a treatment that covers all possible inheritance systems, we have to deal with the possibility, at least in principle, of regressions of offspring on parent that are negative or greater than one In effect, we see that many discussions of heritability have assimilated two different phenomena, parent/offspring *similarity* and parent/offspring *predictability* The former is a special case of the latter

This shows, I think, a slight rupture in our usual picture of the relationships between key theoretical concepts Summaries of Darwinism routinely say that ENS requires that "like must produce like," or that parents must "resemble" offspring [34] But if our aim is to treat the heritability concept as measuring the evolutionary "response" to selection, then we see that there will be some response as long as parent and offspring have *some* systematic relationship, whether this is one of systematic similarity or systematic dissimilarity [35] Cultural inheritance might be a domain where such parent/offspring anti-correlation may be common Different aspects of the theory pull us in different ways with respect to the categorization of such cases They look Darwinian when we are thinking about the abstract idea of a populational response to fitness differences, they look "anti-Darwinian" when we have an eye on the importance of cumulative selection and adaptation

V DRIFT AND CORRELATED RESPONSE

In this section I turn to the most obvious problem with some standard summaries the problem of distinguishing ENS from change occurring via reproductive differences that arise by various kinds of accident Though this problem may initially seem the most difficult to fix, it will receive a comparatively simple treatment here

[34] Endler (*op cit*) is an exception

[35] See Jacquard, *op cit*, and M Blute, "The Evolution of Replication" (forthcoming) Jacquard makes the conceptual separation, and gives separate measures of parent/offspring *resemblance* (k) and parent/offspring *determination* (D), where k is the regression slope and D compares the average variance within offspring classes to the overall variance in the population He then says that, even in a model without mechanistic assumptions, the two will be closely related, and in fact that $D = k^2$ But this result assumes that the variances of the parental and offspring generations are equal, which Jacquard notes but downplays

I organize the discussion with two cases handled simultaneously

Case 8, Accident Individuals' values of Z have no causal role in survival and reproduction But individuals with higher values of Z have more offspring purely by accident. Z is heritable The mean value of Z increases

Case 9, Correlated Response Individuals with higher Z values also have higher values of X. Z is causally inert, but high values of X are advantageous, and individuals with high values of X consequently have more offspring X and Z are both heritable Mean values of X and Z both increase

In Case 9, the problem is what we say about Z Trait X evolves by natural selection, clearly, but change in Z seems to be described by some of the standard summaries as well So Case 9 raises the possibility that a summary needs to be expressed in terms of criteria for evolution *of some particular trait* by natural selection

Of the summaries quoted earlier, Lewontin's are the most susceptible to these problems [36] Both Endler and Ridley make explicit gestures towards ruling out reproductive differences that arise by accident In effect, both require that differences in reproductive output have a *systematic* relation to parental phenotype (clauses E2 and R4) In almost all summaries and sketches of ENS, in fact, one can sense a desire to say *something* that rules out accident, accompanied by uncertainty over the right way to do this Thus we see, as in Endler, Ridley, and perhaps Lewontin in 1970, various kinds of modal or causal loading of the language used to describe fitness differences and their relation to phenotype

A natural first response to these problems, and one that would deal with both pure accident and the hitchhiking phenomena of Case 9, is to add a requirement that there be a *causal link* between the phenotypic variation and the reproductive differences cited in the summary, in order for a given trait to evolve by natural selection Specifically, we might require that the phenotypic variation be partly causally responsible for the differences in reproductive output

The main shortcoming of this approach is that it seems coarse-grained and crude, especially given the large body of theory on the relation between selection and random drift All we have here is a binary distinction either the phenotypic variation played *some* causal role, perhaps a minor one, or it did not One might hope for a more fine-grained treatment

This hope is seen in Okasha's treatment of the issue, which I will take as representative of a family of such approaches (*op cit*, chapter 1) Okasha again draws on the Price equation He notes that in principle

[36] This fact has been noted frequently in discussions

the fitness differences used in the Price equation could be due to chance rather than selection, but argues that the Price formalism can be used to distinguish the roles of the two factors Suppose there is a probability distribution, P, that assigns probabilities of various levels of reproductive success to an organism of a given kind in a given environment Each organism i's realized (absolute) fitness W_i can be expressed as a sum of its *expected* fitness in that environment, W_i^* calculated using P, and a deviation δ_i from that expectation [37]

$$W_i = W_i^* + \delta_i, \tag{7}$$

This breakdown is introduced by Okasha into a Price equation Change in mean phenotype can be represented (ignoring transmission bias) as a sum of $Cov(W^*,Z)$ and $Cov(\delta,Z)$ He claims this "partitions the total change into a component due to selection on Z and a component due to random drift, In principle, that is, if we could discover the probability distribution P, we could determine whether the overall change is the result of chance, natural selection, or a combination of the two" (*op cit*, p 33, and see his equation 1 4) This breakdown also answers recent skeptics about the distinction between selection and drift, Okasha says, because it gives us a common currency, units of Z, in which the contributions of each can be compared [38]

I doubt that the probabilistic breakdown achieves Okasha's goal, however The breakdown in terms of P, if it is available, tells us about the extent to which realized fitnesses conform to expected values But an expected value can itself be produced by accident It can obtain without having a basis in the causal factors that *make* W^* the expected value The degree of match between an expected and an actual value cannot itself tell us *how* the actual value was produced And the question of whether an outcome was due to selection or mere accident is a question of exactly this kind, a question about how a set of realized fitness values were produced

More precisely, if the probability distributions that Okasha uses here were available, then it might be a *necessary* condition for the absence of drift that the realized fitnesses are identical to the expected fitnesses But it is not a sufficient condition

Once this problem with the "deviation from expectation" approach is clear, we seem pushed back towards the simple causal criterion

[37] $W^*_i = \Sigma_j P_{ij} j$ Here the j's are numbers of offspring, and P_{ij} is the probability of i having j offspring

[38] For the skeptical arguments, see Mohan Matthen and Andre Ariew, "Two Ways of Thinking about Fitness and Natural Selection," this JOURNAL, XLIX, 2 (February 2002) 53–83

introduced above I suggest that this is an adequate treatment of both drift and correlated response For evolution by natural selection to act on Z, we require (among other things) that the fitness differences affecting the population have a partial causal basis in variation in Z In the case of drift (Case 7), that requirement is not met because the fitness differences were accidental, at least with respect to the role of Z In the case of correlated response, the association between Z and X may or may not be accidental, but variation in Z played no causal role in the fitness differences that produce change

This shows that the verbal summaries have a very different role in the context of this last family of problems, when compared to their role in earlier sections of this paper Whereas in the earlier sections, the verbal summaries seemed destined to be no more than approximations of more formal treatments, a verbal formulation of the kind discussed here is the most natural and appropriate way of capturing the causal requirement that handles the problems of drift and correlated response This shows also that standard recipes such as Lewontin's were not mistaken to include a pair of conditions, one on heritability and one on the covariance between fitness and parental phenotype, $Cov(Z, w)$ Okasha, as discussed earlier, suggests that the key requirement can be expressed in terms of $\mathrm{Cov}(Z', w)$ alone But moving to a criterion of that kind makes it impossible to capture the causal role of parental phenotype, as it "jumps over" the parental phenotype

Consequently, a uniform treatment of the last three cases (7, 8, and 9) can be given In each case, there is no natural selection on Z because differences in Z had no causal role in producing fitness differences In Case 9, we do have change in the population due to natural selection on X, so there is *some* natural selection Some may think that the standard summaries are only designed to say when there is natural selection present at all, on that view, Case 9 is not a problem In any case, the need to deal with the distinction between selection and random drift (Case 8) is enough on its own to motivate a causal requirement

It might be argued at this point that the problem has not been solved until a more exact specification has been given of the relevant kind of causal link between phenotype and fitness I accept that a more precise treatment of causation would be desirable, in general and in this specific context, but I do not think it is necessary at this point [39] What is needed for present purposes is a way of distinguishing

[39] Here a manipulability approach might be used, see Judea Pearl, *Causality Models, Reasoning, and Inference* (New York Cambridge, 1998), and James Woodward, *Making Things Happen A Theory of Causal Explanation* (New York Oxford, 2003)

a broad and heterogeneous category of changes due to ENS, from changes that might look similar but are due to accident Knowledge of causal facts of the relevant kind might be difficult, but those are the facts that mark the crucial difference Once we know that differences in Z were partly responsible for differences in fitness, we know the case is not a pure case of random drift, and that is the distinction that has to be made for present purposes

So the verbal summaries play a very different role here than they did above In earlier sections, they struggled to accommodate all the cases, and had to be seen as embodying idealizations or approximations In this last section, it is verbal formulations that give us the most natural way (and presently the only available way) of expressing a crucial requirement in a fully general form

VI CONCLUSION

The range of problem cases discussed in this paper is not complete, but is sufficient to motivate some fairly definite conclusions [40]

The project of giving a *summary* of ENS should be kept distinct from the project of giving a *recipe* for change in a population Both projects are worthwhile, but they are somewhat different Problems arise when a single formulation is intended to do both A recipe that is simple and straightforwardly predictive will not capture all cases A summary that covers all cases will not give simple conditions causally sufficient for change

Once these goals are separated, we also see that the procedures and problems faced in each project are different Suppose first that our aim is to give a summary characterization of ENS that will capture all genuine cases Then we are able to *assume* the presence of change, and our aim is to say which changes *count as* ENS Our aim is to mark the boundaries of a category that has a particular explanatory role The decisions that must be made will often concern how broad the category should be, and how we should manage trade-offs between simplicity and scope—the kind of trade-off emphasized in some unificatiomst approaches to explanation [41] This sort of discus-

[40] I have not discussed problems arising from the role of variance in fitness in predicting change, and the occasional need to track grand-offspring rather than offspring when measuring fitness For discussion of both, see Arlew and Lewontin (*op cit*) I should add that I *do not* think the summaries need to be narrowed in order to deal with spurious cases of change at higher or lower levels than the "real" level at which evolution is occurring, but that claim will not be defended here

[41] See Philip Kitcher, "Explanatory Unification and the Causal Structure of the World," in Kitcher and Wesley Salmon, eds , *Minnesota Studies in the Philosophy of Science, Volume XIII* (Minneapolis Minnesota UP, 1989), pp 410–505

sion can also devolve into terminological dispute of the unproductive kind As argued in section II, we should resist the temptation to impose overly sharp borders

So consider the following summary

(S) Evolution by natural selection is change in a population due to

(i) variation in the characteristics of members of the population,
(ii) which causes different rates of reproduction, and
(iii) which is inherited

This summary allows that change may occur because of the inheritance system alone, or by various other mechanisms, ENS is change over and above that resulting from these other factors The reference to "rates" is intended to measure output with respect to time, but this need not be the particular "growth rate" parameter discussed in section II In a related vein, I assume that inheritance involves parent/offspring similarity (as opposed to any systematic relationship), but it is not assumed that a regression is always the appropriate measure of this similarity

This summary does not include within ENS change to due to differences in survival that are not reflected in reproductive differences Those include short-term change due to simple culling, and effects on a population due to some individuals living longer than others after all their reproduction is completed This issue does not arise when discrete generations are assumed But when discrete generations are not assumed, it becomes necessary to work out the relation between change due to survival differences *per se*, and change due to differences in reproduction Ridley's summary rules out change due to survival differences *per se* But as we saw, he makes a tacit idealizing assumption of discrete generations Endler rules these cases in, but does so by giving two descriptions of change, E4 and E5, where E4 handles change due to differential survival and E5 handles the "response" across generations So inheritance matters only to the change in E5 In the summary above I treat only change due to reproductive differences as ENS, but that is a decision of the kind described above—a decision about how broadly to extend the borders of an explanatory category in a short summary, given that there is an admitted grey area

A distinct project is trying to give a *recipe* for change that captures the core features of ENS in a compact and transparent way Now we can help ourselves to idealizations of various kinds—we can assume discrete generations and asexual reproduction, if we want We may also explicitly allow approximation The risk now becomes not a collapse into terminological wrangling, but an embrace of excessive idealizations that lead to a loss of contact with important cases It is also possible to

stop trying to give such recipes in verbal form, trusting instead to a collection of equations (the Price equation, the one-locus diploid model, the replicator dynamics, the breeder's equation) But as emphasized in section v, a verbal summary does a better job with the causal component of the idea of selection than existing formalisms do, and verbal recipes will also be practically useful in contexts where it is necessary to avoid technicality Some recipe-makers may also want to use an equation *plus* a verbal commentary as their preferred form of representation

One idealized recipe can be constructed via modification of Lewontin's 1980 formulation

(IR) The following conditions are sufficient for evolution of trait Z by natural selection in a population with discrete generations

(i) There is variation in Z,
(ii) There is a covariance between Z and the number of offspring left by individuals, where this covariance is partly due to the causal role of Z, and
(iii) The variation is heritable, and inherited without directional bias

Here we indicate a key idealization and rule out problems of inheritance bias explicitly We can assume that heritability is understood in a fitness-weighted way The requirement of covariance between Z and fitness rules out problems with stabilizing selection We could also make the recipe more exact by stipulating an absence of migration Some possibilities might be seen as captured by a tacit *ceteris paribus* clause But once we are in the domain of idealized recipes, there is no need to capture all cases, and no need to settle on a single recipe for all purposes and contexts

Although these idealized recipes do not substitute for a summary of what ENS is, they can be very illuminating They give a compact and causally informative representation of some core cases of ENS, and provide a basis for the analysis of more complex ones Problems only arise when the idealizations that have been made are forgotten or denied

PETER GODFREY-SMITH

Harvard University

AN IMPERATIVE THEORY OF PAIN*

Intentionalism—the thesis that the qualitative properties of an experience are exhausted by its intentional contents—has come as a gospel to many physicalists [1] It promises an end to the hard question of consciousness It opens the way to a satisfying integration of the philosophy and the neuroscience of perception Best of all, there are good philosophical arguments for it

Nevertheless, intentionalism faces difficult cases Many have been drawn towards what Alex Byrne calls *restricted* intentionalism Restricted intentionalism accepts intentionalism for a large class of sensations, but denies it for a special class of bodily sensations—hunger, thirst, itches, and so on The most striking of these problematic sensations are pains

Representational theories of pain have been proposed,[2] but they are problematic First, pains have a distinct qualitative feel that seems to have little to do with their content The pain of a paper cut may be associated with an entirely trivial state of affairs, the pain itself has a strong, nontrivial motivating aspect to it Conversely, cases like that of morphine pain appear to show that motivation can be dissociated from the feel of pain, suggesting to many that ordinary cases of pain are not exhausted by what they report Such cases have pushed many in the field to a dual-aspect theory of pain, on which ordinary pains consist of two dissociable sensory states one associated with tissue damage, and one that motivates (perhaps via an irreducible qualitative feel) towards action [3] In the end, even those who might be

*Thanks to David Hilbert for helpful feedback and discussion on the present version Aderemi Artis, Richard Hall, Sean Kelly, Gabriel Love, Jeff Speaks, Dr Edward Thompson, and audiences at Princeton University and the Association for the Scientific Study of Consciousness (ASSC), 9th annual meeting, also gave invaluable comments on early drafts of this work

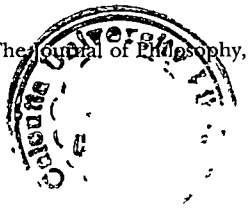
[1] Gilbert Harman, "The Intrinsic Quality of Experience," *Philosophical Perspectives*, IV (1990) 31–52, Alex Byrne, "Intentionalism Defended," *Philosophical Review*, CX, 2 (2001) 199–239

[2] See Michael Tye, "A Representational Theory of Pains and Their Phenomenal Character," *Philosophical Perspectives*, IX (1995) 223–39, Byrne, "Intentionalism Defended", Michael Thau, *Consciousness and Cognition* (New York Oxford, 2002)

[3] For influential arguments along these lines, see Ronald Melzack, *The Puzzle of Pain* (New York Basic Books, 1973), Melzack and Patrick Wall, *The Challenge of Pain* (New York Basic Books, 1983), Valerie Gray Hardcastle, "When a Pain Is Not," this JOURNAL, XCIV, 8 (August 1997) 381–406

0022-362X/07/0410/517–32

© 2007 The Journal of Philosophy, Inc



inclined towards intentionalist theories of pains find it difficult to account for this unique phenomenology This in turn gives fuel to those who are inclined to ditch intentionalism entirely and make a case for primitive pain qualia

Pain presents a problem only for a certain type of intentionalism, however The above critiques assume that the intentional content of sensations must be representational That is, they assume that the content of all sensations is akin to the content of declarative sentences, where the content is something like a proposition or a description This is unnecessarily restrictive Continuing the analogy, there are many types of meaningful sentences, and each has a content Not all sentences are declarative, though, and we are happy with a pluralism of types of content—there are interrogative, imperative, and subjunctive sentences as well as declaratives Each of these sentences has a distinct kind of content irreducible to the others

A similar distinction can save intentionalism about pains I argue that pains are exhausted by their content, but that this content is imperative rather than representational Pains thus command rather than describe Commanding is still a way of having content, however, and so intentionalism is preserved In what follows, I will sketch an imperative theory of pain and defend it against two common objections I suggest that the biological function of pain has been misunderstood, properly conceived, the purpose of pain fits well with an imperative account I conclude with some areas of open research

I AN IMPERATIVE THEORY OF PAIN

I 1 Imperative Sensations Imperative sentences do not describe the world When I insist that a student write a paper, I do not represent a way that the world is I do presuppose that the world is a certain way consistent with her not yet having written a paper But I do not utter my command in order to express the proposition that the world lacks a paper written by my student

The direction of fit of imperatives is rather more like that of desires Both are only satisfied if the world is made to be a certain way Unlike many desires, though, imperatives are not just satisfied by the world coming to be a certain way My command to a student is not satisfied by the production of a paper with my student's name on it, instead, in issuing imperatives I (explicitly or implicitly) demand a certain action from my student herself My imperative is thus satisfied if and only if the target of the command acts in a certain way, my command to my student is satisfied by her activity of paper-writing, not by the written paper itself

Imperative sentences demand an action, similarly so with imperative sensations Richard Hall has argued[4] that the content of an itch is the command 'Scratch here!', the content of hunger is 'Eat Something!', and so on These sensations obey a logic similar to that of imperative sentences First, they do not represent the world as being a certain way When I feel an itch, I do not know what to expect at its location There are many causes of itches, and the sensation itself does not tell me which of them might obtain Further, an itch is not the sort of thing that can be veridical or nonveridical Since itches are not in the business of reporting, we are not surprised if we cannot even find a reason why we itch This suggests that itches are not usefully understood as representing anything

What an itch does do is demand that I perform a certain action—namely, scratch it Again, this is not equivalent to the passive claim that the world be made a certain way There is no particular state of affairs that an itch demands as an itch does not imply that the world is a certain way, it does not require it to be made any other way Instead, my itch commands only a certain action from me, that of scratching I might satisfy this imperative in a number of different ways scratching myself with my nails, or rubbing up against a rough tree, or fishing around with a ruler What is demanded by an itch is unified only by the action that it demands—that is, by scratching Thus, I do not perceive an itch as a passive subject who merely wants the world to be a certain way I approach it as someone who must do something to satisfy a command issued by my body

Other imperative sensations include hunger and thirst Hunger demands that I eat, thirst demands that I drink Note the features common to all of these sensations Each carries little information about either its cause or about what future state of the world would be required to alleviate it Indeed, all display a curious failure even to localize precisely, and some (like thirst) can be felt as imperatives that issue from nowhere in particular at all The distinctive qualities of each stem from the fact that they demand a certain kind of action to be satisfied

I 2 Pains as Negative Imperatives Pains have the same features as imperative sensations They are often poorly localized, they distract from action and (as I will argue in section II) they do not carry information about their causal origins[5] This is good evidence that

[4] In, alas, still unpublished work Hall has indicated (in private correspondence) that he thinks pains are viable candidates for an imperative treatment. However, none of what follows should be considered to express Hall's considered views except where noted

[5] For the relation between pain and itch, see M Schmelz and H Handwerker, "Itch," in Stephen B McMahon and Martin Koltzenburg, eds , *Wall and Melzack's Textbook of*

pains are also imperatives they are felt as commands, and impinge upon the mental economy as do other imperatives

There is a small obstacle to treating pains as imperatives Other imperative sensations are unified by a single action type that would satisfy them (eating, scratching, and so on) Most pains do not seem to demand any particular positive course of action It is true that some acute pains cause withdrawal, but withdrawal can be explained by spinal reflex When we consider the effect of pain on planned actions, we find no simple action type that would satisfy them Consider in this regard the pain in my thighs as I run up a steep hill, or the ache in my arms as I repeatedly lift a heavy weight, or the pain in my ankle as I recuperate from a sprain None of these demand that I do anything in particular

Phrased this way, however, I think the solution is clear The content of any pain is a negative imperative [6] The imperative force of pains is thus to proscribe rather than prescribe What unifies the above pains is the imperative that I stop doing what I am doing, their content is a proscription against action

The more specific content of pain depends on where and how it is felt Take the pain of a broken ankle The pain of a broken ankle is felt as a command against putting weight on that ankle The content of that pain is therefore a negative imperative against moving in a way that would put weight on the ankle There is no positive content

---

Pain (Philadelphia Elsevier, 2006), pp 219–27 For further parallels with bodily sensations, see Wall, "On the Relation of Injury to Pain," *Pain*, VI (1979) 253–64, especially p 254

[6] Other negative imperatives might include nausea and the various nameless aches associated with illness I leave it open what might differentiate them from pains, but I do assume that there is something in their content that distinguishes them (unlike Nathan Nelkin in his "Reconsidering Pain," *Philosophical Psychology*, VII (1994) 325–43)

I also assume that there is something that separates *sensations* with a negative imperative content from, for example, strong *thoughts* with negative imperative content What this difference might be is an interesting problem, and one that faces representationalist forms of intentionalism Representationalist responses have been well discussed in the literature, for a good recent review of some distinctions between styles of response, see section 2 of Andy Egan, "Appearance Properties?" *Noûs*, XL 3 (2006) 495–521 I intend the present account is compatible with any of these proposed solutions It could be, for example, that thoughts with negative imperative content are distinguished from sensations with negative imperative content by their functional roles (what Ned Block calls 'quasi-intentionalism' in Block, "Mental Paint and Mental Latex," *Philosophical Issues*, VII (1996) 19–49) It is also possible that the imperative content of pain is sufficiently unique to distinguish it from other sorts of imperatives I rather suspect that the latter is the case the imperatives of pain proscribe against a wide range of complex ways of bodily action, and so differ from thought imperatives just in virtue of their additional content

required as long as one avoids certain ways of acting, the imperative of pain will be satisfied

The account extends to other pains The pain of a burn commands against any action that would cause the injured area to contact the world As such, pain encourages us to treat burns (and scrapes and wounds) gingerly The pain of a sprained joint is a negative imperative against moving the limb along one or more of its degrees of freedom The pain of incipient damage—say, as I bring my hand close to a fire—is an imperative against any action that would bring my hand even closer The pains of exertion and exercise are felt imperatives against continued use of the affected body part

The imperative account also has a story about the content of pains usually associated with reflexive withdrawal Withdrawal because of pain may seem problematic, for it looks like a case where the relevant imperative would be positive rather than negative Not so The imperative account treats these as limit cases The pain of touching a hot stove is a very strong imperative against any action that would keep your hand in contact with the stove (including the trivial action of keeping it exactly where it is) Some sort of withdrawal is thus effectively demanded However, it is demanded only indirectly, via a command against doing anything but withdrawal

A few quick notes are in order First, I claim that pains attach to the use of body parts, rather than just to body parts themselves In this regard, I think the pain literature is overly fond of a problematic locution Many philosophers take as paradigmatic statements like 'My *B* hurts' or 'I have a pain in my *B*', where *B* stands for a particular body part If I am right, however, such locutions are misleading—not in the least because they encourage one to ignore a large part of the content of pains Instead, the telling locution is 'It hurts when I *A*' or 'My *B* hurts when I *A*', where *A* is an action

Of course, this is not to deny that pains are felt as located in parts of the body Further, there is surely an intimate relation between the felt location of pains and the actions proscribed This should be expected Again, pains do not proscribe against particular actions directly In many cases there is an indefinite number of actions against which a pain would have to proscribe There are many things that you should not do when you have a broken ankle (standing, kickball, hopscotch) The content of pain certainly does not range over all of these These are all effectively ruled out as a consequence of the general imperative against using your body in ways that would further injure your ankle by putting weight on it

A second point I have spoken as if we perceived pains while static and idle This is a simplification When we feel pain, we are usually

moving in the world, executing planned actions to achieve 'goals Pains show up against this background of planned action as imperatives against performing certain actions in various ways The imperatives of pains thus enter into a larger process of planning action, and can be overridden within that context I can choose not to listen to the pain caused by the hot casserole in my hands, at least for long as it takes to get the dish to the table Pains are dynamic, however, and we ignore the pains of action only for a time Pain tends to increase as we perform a prohibited action, and decrease as we back away Hence the strength of the imperative can differ as we go about the world, exerting a correspondingly different effect on our action as it does

Finally, I argue that the imperative content of pains exhausts what there is to say about pain This is meant in support of the intentionalist The phenomenal properties of pains are exhausted by their imperative content, and nothing more need be cited to account for the qualitative feel of pains The difference between someone in pain and someone not in pain is just the presence or absence of an imperative The pain of a broken ankle is an imperative for me not to use my ankle Give me enough morphine for my pain to be lessened, and you lessen the strength of that imperative Give me enough to eliminate my pain, and you have eliminated the imperative against action entirely Nothing more need be postulated to explain my phenomenology, nor to explain the distinctive contribution that phenomenology makes to my behavior Intentionalism in its broadest form is thus true of pains there is no difference in the phenomenology of a pain experience without some difference in the intentional content of a pain experience

II OBJECTION 1 TISSUE DAMAGE

So far I have said nothing about tissue damage—and, in particular, the role that pains might play in informing us about such damage Many would object to this For on most accounts, we feel pain precisely because we need to be informed about tissue damage [7] This is often backed by a seductive argument the biological purpose of pains is to inform about damage, and so the content of pains must have some informative content commensurate with this role [8] Call this view the *myth of tissue damage* The imperative account says nothing about

[7] Or any similar state—I will use tissue damage as a generic term to includes disturbance, disorder, or any other harm to the body

[8] This view is not confined to representational accounts, something like it is also assumed by many qualia theorists and those with a more complex view of the phenomenology of pain See, for example, Hall, "Are Pains Necessarily Unpleasant?"

tissue damage This may seem like a problematic lacuna However, the omission is deliberate The myth of tissue damage is false, and pains are only incidentally related to tissue damage

Before the arguments, an important point We frequently come to know that we have been damaged because we feel pain I do not deny that obvious fact But that alone does not weigh against the imperative account A soldier may learn that war has begun by receiving his marching orders That does not show that the content of marching orders includes information about the commencement of hostilities Nor does it show that his marching orders were issued in order to inform him about some state of affairs The content of his marching orders is entirely exhausted by the imperative they contain So, too, with pains Pain, like any imperative, may indirectly inform us about the occurrence of the situation that caused it to be issued

With that in mind, there are three reasons to deny that the myth of tissue damage First, pain is often caused by states other than actual tissue damage Second, pain is not a reliable source of information about tissue damage Third, the biological purpose of pain is not to inform, and would be undermined by sensations with informative content Let us look at each of these points in turn

First, tissue damage is not the sole, or even typical, cause of pain Many pains show up in nondamage situations We feel pain after acute damage But we also feel pain before damage—as when we get our hands too close to a flame This pain does not report on damage, because no damage has yet occurred, indeed, if we listen to our pain, no damage will occur We feel chronic pains following injuries like broken bones These pains persist during recuperation and long after the original damage has healed Hence, these pains are associated only with weakness and propensity to re-injury We also feel pain during heavy exertion, in many cases, exertion is not even counterfactually connected with damage, as we would collapse from lack of energy before we could hurt ourselves

Importantly, none of these are pathological cases, failure to feel pain in these circumstances would put one at serious disadvantage This is most dramatically shown by those congenitally insensitive to pain Their unfortunate lives are typically cut short not by acute damage, but by the collective impact on their joints of re-exacerbated

---

Philosophy and Phenomenological Research, XXIX (1989) 643–49, Murat Aydede and Güven Güzeldere, "Some Foundational Problems in the Scientific Study of Pain," *Philosophy of Science*, LXIX (2002) S265–S283

injuries and of long-term overexertion [9] The pains associated with potential injury, exertion, and postural adjustment are thus just as important for our survival as the pains associated with acute damage

Further, we mostly learn to withdraw from pain before tissue damage Similarly, the pains of exertion are common, and the pains of injury-weakened tissues last for months Hence, these nondamage pains may well account for the majority of the pains the average person feels The assumed correlation between pain and tissue damage thus begins to look like a failure of imagination We feel pains for a heterogeneous variety of bodily causes, few of which involve damage

Second, those in pain are often poorly informed about its cause Given that pain is so poorly correlated with tissue damage, this should come as no surprise To take an obvious case, even with the help of advanced diagnostic tests only about 10% of cases of lower back pain can be assigned a specific origin [10] Lower back pain thus carries little indication about whether one's lower back is damaged, diseased, or abnormal Visceral pains are similarly ambiguous, patients experiencing acute abdominal pain are often unable even to localize the source of the pain with any reasonable precision [11]

Nonclinical cases are also common Consider it remains an area of active research why muscles hurt after long labor—and in particular whether the cause is microdamage to muscles or some other nondamage mechanism Pain itself does not appear to give those so suffering any information to help resolve this question, nor do athletes usually treat such pains as indicators of damage [12] Similarly, anyone as clumsy as the author will have had the experience of running into a large object, assuming that one has been injured, and discovering a few minutes later that one is quite intact and unharmed Pains can thus be a poor source of information as to their origin—puzzling if the content or typical cause of pains is supposed to be tissue damage [13]

[9] And barring that, they will succumb to osteomyelitis caused by wear on joint edges This is not a result of acute injury, it stems from a failure properly to regulate posture (Melzack and Wall, *The Challenge of Pain*, pp 18–19) Those who contemplate the relation between pain and action would do well to reflect on the vital, usually overlooked role played by pains in this regard

[10] M van Tulder and B Koes, "Low Back Pain," in McMahon and Koltzenburg, eds , *Wall and Melzack's Textbook of Pain*, pp 699–708, p 699

[11] H Y Wong and E A Mayer, "A Clinical Perspective on Abdominal Pain," in McMahon and Koltzenburg, eds , *Wall and Melzack's Textbook of Pain*, pp 753–76, p 759

[12] U Proske, "Muscle Tenderness from Exercise Mechanisms?" *The Journal of Physiology*, DLXIV (2005) 1

[13] Conversely, many serious injuries are in fact painless for a time This is also difficult to reconcile with views that suppose that the purpose of pain is informative, since it appears that pain fails to show up precisely where it should Further, such cases are

Examining a sufficiently broad range of cases therefore casts doubt on any view that treats the content of pains as a report on tissue damage [14] Of course, it is not impossible to maintain that the content of pain is nevertheless tissue damage, I take it, however, that these cases should make that view increasingly implausible

Third, the biological function of pain is not plausibly interpreted as informing about damage The biological role of pain clearly has something to do with keeping bodies healthy and intact Those unlucky enough to be born congenitally insensitive to pain again offer dramatic proof of the utility of pain The interesting question is how pain protects us, and whether that protective role would best be served by a sensation that informed us when damage occurs

To begin, note that the cases where pain keeps our bodies healthy are all cases where motion needs to be limited for various reasons—to prevent future injury, or further injury, or re-injury, to name a few [15]

---

not easily explained by distraction, pathology, or top-down cognitive control See H Beecher, "Relationship of Significance of Wound to Pain Experienced," *Journal of the American Medical Association,* CLXI (1956) 1609–13, Beecher, *Measurement of Subjective Responses to Pain* (New York Oxford, 1959), P Carlen, Wall, H Nadvorna, and T Steinbach, "Phantom Limbs and Related Phenomena in Recent Traumatic Amputations," *Neurology,* XXVIII (1978) 211–17 For compelling evidence against positive cognitive evaluation of injury as the explanation of painless injury, see Melzack, Wall, and T Ty, "Acute Pain in an Emergency Clinic Latency of Onset and Descriptor Patterns Related to Different Injuries," *Pain,* XIV (1982) 33–43 What follows owes a debt in part to Wall's discussion of painless injury in the course of developing his multi-stage theory of pain, especially in his *Pain* (New York Columbia, 2000)

[14] The required unity will not come from appeal to lower-level physiological features like pain receptors Melzack and Wall (*The Challenge of Pain,* especially chs 5–7) provide a useful overview of the physiology of pain perception, and it is clear that nothing like a straightforward physical pathway for pain exists Not all pain has its source at specialized receptors, there is a good deal of neural processing even at the spinal level (including a complex system of peripheral and central inhibition), and there are probably several spinal pathways for nocioceptive information

[15] As Wall emphasizes, there is a strong neurological link between pain and some sort of planning for action Functional imaging studies show that the ventral premotor cortex, an area involved in planning motor actions, is strongly and consistently activated by pain R Coghill, C Sang, J Maisog, and M Iadarola, "Pain Intensity Processing within the Human Brain A Bilateral, Distributed Mechanism," *Journal of Neurophysiology,* LXXXII (1999) 1934–43 The same area contains premotor neurons that have dual receptive fields—both a tactile and bodily- (rather than retinally-) centered visual one, suggesting a role in integrating touch and sight for future action—Michael Graziano, Gregory Yap, and Charles Gross, "Coding of Visual Space by Premotor Neurons," *Science,* CCLXVI (1994) 1054–57 Furthermore, there is good evidence that pain activates numerous regions of the anterior cingulate cortex, which is involved in attention, motivation, and motor planning Donald Price, "Psychological and Neural Mechanisms of the Affective Dimension of Pain," *Science,* CCLXXXVIII (2000) 1769–72, Pierre Rainville, Gary Duncan, Price, Benoit Carrier, and M Catherine Bushnell, "Pain Affect

Now suppose that the content of pains was information about tissue damage (or, more broadly, some other state that required limitation of motion) On such a view, pain would be like the sensation of blue, or the smell of a magnolia—that is, it would inform us that a particular state of affairs obtained

Therein lies the problem The right action to perform when we see blue is highly variable It depends on the context, on what we know about the blue thing, and so on For many sensations, this contingency of response is a functional advantage it allows for flexible response to the same property in the external world This is clearly an advantage in complex environments

Contingency of response would be a positive disadvantage when it comes to pain, however In cases where we feel pain, there is only one appropriate response, other things being equal That is, to cease acting in a way that would cause more injury A merely contingent relationship between the content of pains and the biologically crucial response of limiting movements would therefore only hurt us, at least in the vast majority of cases As such, an informative content to pain would be positively maladaptive Far from supporting the myth of tissue damage, therefore, considering the biological role of pain weighs against it [16]

Contrast that failure with the imperative account On an imperative view, the primary role of pain is to limit action by commanding against using your body in certain ways This is a good way to achieve the biological end of feeling pain Given the wide range of circumstances and the importance of limiting motion, pains should not have the luxury of informing or persuading Their important function instead requires that they command against movements contrary to health This is precisely what the imperative account claims

The imperative sensations generally have the feature that informative content (and the resulting possibility of contingency of response) would be maladaptive An animal who has not had food or drink in days does not need to be informed about her body She needs to be driven to eat and drink, a biologically effective system for doing so should not bother with the niceties of explaining why So, too, with pain

---

Encoded in Human Anterior Cingulate but Not Somatosensory Cortex," *Science*, CCLXXVII (1997) 968–71, O Devinsky, M Morrell, and B Vogt, "Contributions of Anterior Cingulate Cortex to Behavior," *Brain*, CXVIII (1995) 279–306

[16] Thanks to David Hilbert for helpful discussion on this point

III OBJECTION 2 MORPHINE PAIN

The imperative theory belongs to a group we might call motivational theories of pain [17] According to motivational theories, feeling pain has some essentially motivating character The imperative theory cashes that character out as a sensory imperative against using your body in certain ways

Motivational theories of pain often founder on the problem of morphine pain Some patients, when given morphine shortly after acute trauma, will report that their pain persists, but that they no longer care about it They do not seem motivated to do anything about their pain—complain, ask for help—and they seem quite content Morphine pain is not a rare fluke, either, most emergency room doctors are familiar with it, and a similar phenomenon occurs with several other types of drugs [18]

In cases of morphine pain, patients therefore report that their pain persists, but that it no longer bothers them—that they do not want to do anything about it This is why many authors have concluded that pains only contingently motivate, since motivational import can be divorced from pain, it is at best an inessential aspect of the sensation [19]

[17] Other motivational accounts include Nelkin, "Reconsidering Pain", B Helm, "Felt Evaluations A Theory of Pleasure and Pain," *American Philosophical Quarterly*, XXXIX (2002) 13–31 I also include Wall's action plan theory of pain, sketched in his *Pain* See Melzack, *The Puzzle of Pain*, pp 147–48, for some historical theories in this camp

[18] There are reports of similar phenomena caused by synthetic inu agonist opiods (such as Meperidine) that differ structurally from morphine, various barbiturates (A Keats and Beecher, "Pain Relief with Hypnotic Doses of Barbiturates and a Hypothesis," *Journal of Pharmacology and Experimental Therapeutics*, C, 1 (1950) 1–13), and nitrous oxide (Hall, "Are Pains Necessarily Unpleasant?") This suggests that the phenomenon does not depend on some effect specific to morphine

[19] See Beecher, "Relationship of Significance of Wound to Pain Experienced", Hall, "Are Pains Necessarily Unpleasant?", Tye, "A Representational Theory of Pains and Their Phenomenal Character", Aydede and Güzdeldere, "Some Foundational Problems in the Scientific Study of Pain", Daniel Dennett, "Why You Can't Make a Computer That Feels Pain," in *Brainstorms* (Cambridge MIT, 1978), pp 190–229 Also tempted by this line, though they waver, are George Pitcher, "The Awfulness of Pain," this JOURNAL, LXVII, 14 (July 23, 1970) 481–92, and Barry Stimmel, *Pain, Analgesia, and Addiction* (New York Raven, 1983)

Price has argued, on roughly these grounds, that the processing is linear, with the affect normally added only after sensory discrimination Further, Coghill et al have presented evidence that pain processing is parallel and partially redundant, a view that probably best supports my own position

Hardcastle (in "When a Pain Is Not") uses the apparent dissociation of pain and unpleasantness to argue that pain research represents a classic use of the double dissociation methodology in psychology—that is, either pain affect or sensory pain can occur on their own The evidence that the primary affect of pain can occur without any sensory aspect, however, relies on ambiguous experiments involving electrical shocks to exposed tooth pulp under the influence of Fentanyl (Richard Gracely, Ronald Dubner, and Patricia McGrath, "Narcotic Analgesia Fentanyl Reduces the

I have argued that some sort of motivation is not just necessarily connected to pain but actually constitutive of it The imperative account therefore faces this objection It is an advantage of the account that it can do so

First, an important point Patients say a lot of things when they report on morphine pain, and it is easy to get distracted by the details Many have supposed that they are meant to explain near-to-actually paradoxical sensations described as 'pain that does not hurt' or 'pain that is not painful' For what it is worth, these appear to be less common than phrasings like 'It hurts but I no longer care' [20] However, philosophers would do well not to get wrapped up in particular turns of phrase [21] Those reporting on morphine pain have gone through recent trauma and are under the influence of powerful narcotics, they are far from the armchair and the phenomenological precision we philosophers would prefer The main thing to explain in morphine pain is how patients can be in severe pain yet not be motivated to do anything by it Particular utterances can be accommodated within an adequate framework as needed

Second, a distinction We have not dealt with the motivational side of pain precisely enough In discussions of morphine pain, the primary motivator is usually taken to be something like the "immediate unpleasantness" of pain Following Price's terminology (though not his exact distinction), I will call this the *primary affect* of pain [22] On the imperative account, primary affect is indeed a central part of pain, and stems from the imperative nature of pain itself Contrast this with *secondary affect*, which includes emotional responses evoked by pain—especially second-order responses to the sensation of pain itself

---

Intensity but Not the Unpleasantness of Painful Tooth Pulp Sensations," *Science*, CCIII (1979) 1261–63) In this case, it seems reasonable to assume that having one's tooth pulp electrically shocked is just unpleasant simpliciter Indeed, I find it hard to see how we would go about distinguishing "pain with the unpleasantness but without the sensory qualities" from simple sensations of unpleasantness

[20] Thanks to Dr Edward Thompson for valuable, if admittedly anecdotal, evidence in this regard

[21] Compare to masochism Philosophers often go astray when they take seriously the simple formulation 'pain that feels good' This is clearly too simple a formulation of what the masochist feels As Pitcher has pointed out, if the whippings felt good, then the masochist would seek out a new partner—his goal, in an important sense, is to be hurt (Pitcher, "The Awfulness of Pain," p 485) Furthermore, as Robert Stoller has emphasized, masochism cannot be separated from the overall context in which it occurs, suggesting that the pain is sought not because of a simple inverse valence but because of its relation to a larger goal of humiliation, sexual excitement, and so on (Stoller, *Pain and Passion A Psychoanalyst Explores the World of S&M* (New York Plenum, 1991), especially ch 1)

[22] Price, "Psychological and Neural Mechanisms of the Affective Dimension of Pain "

Secondary affect includes anxiety about the present and future, fear about what pain signals, and so on

Secondary affect certainly promotes positive activity, often of a complex and intentional kind And morphine's effect is no doubt partially due to an elimination of negative secondary affect But no one should doubt that secondary affect is separable from primary affect Indeed, it has such a contingent connection to token pain experiences that is easier to treat secondary affect as an evoked reaction rather than an inherent part of the pain experience itself As such, arguments for separability of secondary affect do not touch questions about the essential motivational import of pain

Morphine pain is problematic only if it eliminates primary affect If a motivational theory of pain claims that the primary motivation is positive, then that theory does have a problem morphine pain does not appear to motivate patients to do anything at all However, the imperative theory makes no such claims It implies only negative primary affect, pains only proscribe

Patients acutely dosed with morphine are profoundly apathetic and (in emergency room situations) immobile [23] Observed cases of morphine pain have thus occurred in patients who have temporarily lost their drive to do anything, further, the environment they are in demands nothing from them

Nothing prevents pains from being experienced in this state But in this atypical context, pain is completely irrelevant there is no potential activity against which its proscriptions might have force Morphine pain is like a stop sign in a ghost town it possesses all of its negative imperative force, but with no one around to care It is thus no puzzle why sedated patients in these odd cases are not bothered by their pain Since pains do not necessarily motivate new, separate actions, an imperative role can be fulfilled without conflict Hence we have an explanation of the odd phenomenology of morphine pain patients are in pain, but do not care about it because they do not care about doing anything Pain is still present, and just as painful as before, it has simply ceased to be something to care about

This gives a handy prediction The imperative theory predicts that if patients in the grip of morphine pain were to act, then their pain would still motivate against using their bodies in certain ways Emergency rooms present certain practical and ethical hurdles to experimentation, but there is another place where this prediction can be tested

[23] See Robert M Julien, *A Primer of Drug Action* (New York W H Freeman, 1998, 8th ed), ch 10

An identical phenomenon occurs in lobotomized patients [24] Rare now, lobotomized patients were once studied extensively They make reports identical to the reports of those in the grip of morphine pain They too seem profoundly unmotivated, caring little about their pains (or anything else) As it turns out, when they do act, their activity is precisely as the imperative theory predicts Melzack reports that these patients still withdraw from pinprick, avoid walking on broken ankles, and rubbing burned skin [25]

Such patients are thus still motivated by the proscriptive force that I have argued is constitutive of pain They lack secondary affect, and so are not motivated to do anything positive about their pain, nor do they seem to mind too much when they are in pain When such patients do act, however, their actions show that the negative imperative of pain is at work in their mental economy They avoid problematic activities accordingly Thus, we can conclude that the imperative theory of pain, far from being threatened by morphine pain, handles it well

IV REMAINING PROBLEMS

I have sketched an imperative theory of pain and showed how it can overcome certain objections I find the theory convincing in its outlines, but there is still work to be done to develop a full theory

Some of this work will involve dealing with aspects of pain I have passed over This includes pains that do not seem to fit well with the account sketched above, because they do not appear to have the tight connection to action that most pains have Among these are the pains associated with headache, menstrual cramps, and disturbances of the deep viscera [26] It is difficult to see which actions these pains weigh against At best, they proscribe against large movements of the head or torso At worst, the imperative theory must treat them as degenerate cases cases where pain is proscribing against unspecified or unintelligible actions I do believe that they can be accounted for along the lines of the former, but more work would need to be done to tell that story

[24] Indeed, one occasionally gets references to morphine in this context as a "pharmacological lobotomy" See Keats and Beecher, "Pain Relief with Hypnotic Doses of Barbiturates and a Hypothesis"

[25] See Melzack and Wall, *The Challenge of Pain,* p 131 Note that in Melzack and Wall's claim that "The predominant effect of lobotomy appears to be on the motivational-affective dimension of the whole pain experience," they clearly mean secondary affect

[26] These are arguably problems for every account, but a problem for all is a problem for each

I have also said little about the qualitative differences between pains There is no doubt that such differences exist—some pains are stabbing, others burning, and so on It is sometimes suggested that these differences correspond to differences in the type of injury that provokes them, but this turns out to be empirically false [27] I suspect that they may correlate better with more specific types of action being proscribed, that is an empirical question, however, and one where more work is required before a definitive answer can be given

A further aspect of pain that I have omitted is its relationship to what is termed psychological or emotional pain I know of few philosophical accounts that attempt such a link [28] The phenomenal linkage between these pains and physical pains is obvious for many people, however, and a philosophical account entirely ignores them at its peril Again, I find the link between pains and limitation of action to be suggestive in this case, again, I have little more to offer than suggestion

The final aspect of pain that I have not touched is the intimate connection between pain and attention Pain captures attention, and attention modulates pain [29] Both directions are clearly important for our experience An adequate account of phenomena like hypnotic analgesia will certainly require a story about the connection Here I assume that the imperative view probably has plenty to say, for any adequate account of imperative sensations should account for their general ability to capture attention, pain is no different in this regard than thirst or hunger or nausea The relationship between pain and attention will thus likely depend on a worked-out theory of the general relationship between imperative sensations and attention

Despite these outstanding questions, I take it that I have shown the viability of an imperative theory of pain and the problems with alternative accounts Other accounts of pain go wrong because they assume that the body is just another object out in the world that we need to be kept informed about That is wrong The body is our means of getting around and interacting in the world Bodies are made with the capability of injuring themselves through their own

[27] For an example of such a suggestion, see Tye, "A Representational Theory of Pains and Their Phenomenal Character," p 235 For empirical work against this thesis, see Melzack, Wall, and Ty, "Acute Pain in an Emergency Clinic Latency of Onset and Descriptor Patterns Related to Different Injuries," and Melzack and Katz, "Pain Assessment in Adult Patients," in McMahon and Koltzenburg, eds , *Wall and Melzack's Textbook of Pain*, pp 291–304

[28] For one, see Helm, "Felt Evaluations A Theory of Pleasure and Pain "

[29] See C Chapman, "Pain The Perception of Noxious Events," in Richard Sternbach, ed , *The Psychology of Pain* (New York Raven, 1978), pp 96–202, esp pp 179–88

BOOK REVIEWS

In the Beginning Was the Deed Realism and Moralism in Political Argument BERNARD WILLIAMS Selected, edited, and with an introduction by GEOFFREY HAWTHORN Princeton Princeton University Press, 2005 xx + 174 p Cloth $29 95

There is a common refrain that runs through these thirteen lapidary, elegant reflections by the late Bernard Williams It is that some sorts of decisions, or disputes, or actions, or expressions of value, are best considered as primarily "political," where this in most cases means not "moral" or not "based on moral principles " The key difference concerns the nature of disagreements For one thing, the uniqueness of moral discussions derives from the nature of the reasons adduced in such disagreements A moral argument seeks to classify an action or practice as permissible, forbidden, or obligatory In political contexts disagreements are meant to have special sorts of consequences actions by governments, institutions that possess a monopoly of legitimately applicable coercive force, and these always occur in particular circumstances that resist generalization For another, the loser in a moral argument is judged to be morally wrong or immoral, in a political dispute she has simply lost a contest, one of many that will arise and pass away

Williams of course realizes that moral considerations play a role in politics, he just does not want them playing the leading or trumping role That is, the fact that people are suffering is certainly a proper and moral consideration in political discussions, but the point Williams is making is that we need to understand how the bearing of that or any other moral concern changes a very great deal when it is introduced as a reason for some government to do something This is primarily because anything such a government can do unavoidably constrains, with the backing of that monopoly on the use of force, what others would otherwise be able to do Williams's claim is that this latter fact changes everything The fact that someone or some group is doing something immoral does not entail that I or anyone ought to, is entitled to, or much less must, stop them An additional and different reason is necessary to justify that claim

Moreover, political disputes assume a "legitimation story" common to both parties and necessarily accepted if the losers are to concede that they lost and nonetheless submit to government power It is an important aspect of Williams's position that this legitimation question

0022-362X/07/0410/533–39 © 2007 The Journal of Philosophy, Inc

common to all human attempts to distinguish the mere exercise of power from genuinely political claims is not itself a moral requirement, or at least not one "prior" to politics It is rather "inherent in there being such a things as politics" (5) It may be that in some sense coercing people to do what one wants is immoral, and that thus persons can be said to have a moral right, to be morally entitled, to be free from such domination, but Williams objects forcefully and often against what he sees as a typically American attempt to consider everything that "ought not to happen" as correlated with a person's "politically enforceable right" not to have it happen This, he says, distorts the true, historically inflected nature of the political and poisons political debate with a rigid moralism (64)

By contrast, moral views of the political hold either that philosophy or human reason can formulate ideals and values which then must be translated into political idiom and "enacted " (He claims that utilitarians often think this way) Or philosophers formulate minimum structural constraints for any political institutions what, minimally, may never be ignored, must always be attended to, in any institution or practice, all perhaps based on some ideal of persons as free and equal rational beings (Williams thinks Kantians and Rawlseans argue this way They do not think of politics as a *modus vivendi*, but they attend to these structural conditions with our "highest moral powers, above all a sense of fairness" in mind (2))

By contrast, Williams's position on how to reason normatively in the political domain is more modest and skeptical He works back, in effect, from the starkest political question Under what conditions, in consideration of what ends, is coercing people—in effect, threatening them with violence—acceptable? He means acceptable really, not just in principle, to those affected (all subject to the "critical theory" worry that this cannot simply be a matter of actual, empirical acceptance), and by conditions he means historical conditions That is, these legitimation stories change across historical time The most important change has been that many ambitious claims to legitimacy (inherited privilege, natural titles to rule, divine right) are now simply no longer credible So distinctly political disputes operate under strict constraints about what "is historically acceptable and so possible" at a time Accordingly, for Williams one of the greatest weaknesses in liberalism is that it has "no account of the cognitive status of its own history," even though its basic superiority consists in its unique compatibility with current historical conditions (One *could* imagine "Kant at the Court of King Arthur," wagging his finger at their injustices and oppression—and Williams hints that he thinks Kantians do like to imagine such things and make such points—but such exercises are for him pointless and self-indulgent (66))

The massive historical constraint on our own political disputes is, in a word (his word), "modernity" So his "modest" formula is the basic legitimation demand + modernity = liberalism, where that roughly means constraints on liberty only to secure a like liberty for all, and a basic solution, compatible with that constraint, to the problems of security and the conditions of trust and cooperation Acknowledging all of this amounts to acknowledging what Williams calls "political realism "

Whether Kant or that Kantians are subject to these charges will be a matter of some contention Kant after all had his own theory of history and of the Enlightenment (though it is not quite the "error theory" Williams demands—why complex human civilizations have only come to realize the superordinate value of liberty after a few thousand years[1]), and his *Metaphysics of Morals* strictly differentiates between the concerns of morality and those of political thought along lines informed by concerns similar to Williams (the justifiable use of coercion) And there are certainly Kantians who are willing to consider who we have historically come to be as indispensable to moral reflection [2] There is also the question of how one would want to distinguish between some dreadful and degraded historical conditions which leave only certain political options open, and on the other hand those constraints on possible legitimation strategies which one wants to affirm besides merely accept Thinking of the matter in this way could give one pause about just how "modest" such a formula is Some progressivist claim seems necessary to make this distinction It may be appropriate to regard "the discovery of what rights people have as a political and historical one, not a philosophical one," (61) but the claim that this *is* a "discovery" is surely a philosophical one

Williams's is a recognizable position in the history of political thought, widely construed He is aware of the resonances of his position in figures such as Thucydides, Machiavelli and one of the most important for his later work, Nietzsche, and his worries about fanaticism, rigidity, and the substitution of moralism for "sound political judgment" are common to writers like Isaiah Berlin and Michael Oakeshott (Williams prefers to distinguish the "liberalism of natural rights" (the Lockean liberalism of "energetic litigants") from the "liberalism of personal development" (Mill), from his preferred "liberalism of fear" and skepticism as formulated by Judith Shklar and visible in Montaigne and Constant) The distinctiveness of Williams's

[1] Kant's unusual answer is not an explanation but a vague indictment What accounts for this fact is a kind of cowardice, a "self-incurred" immaturity

[2] Cf Christine Korsgaard, *The Sources of Normativity* (New York. Cambridge, 1996), p 117

own position consists in (a) the fine-grained accounts of what such a realist position would have to say about such issues as toleration, humanitarian intervention, and censorship and (b) his ambitious treatment, in the realist mode, of the primary modern political value, liberty These then are the three dimensions of the book the underlying opposition between realism and a moralized view of politics, the discussion of liberty "as a political value" and its putative contest with the value of equality, and the application of his realism to contemporary political debates

It is useful to consider these latter essays first, especially since Williams is himself often a proponent of a "bottom up" rather than "top down" methodology That is, one can see what is at stake for him in framing disputes as political Consider the question of "the right (or even the duty) to intervene" for "humanitarian purposes " Much of our reasoning here, Williams argues, is based on an analogy with the morality (rights and duties) of rescue But politically what is often at issue is that *X is in peril at the hands of Z* and this means, contrary to most rescue cases, that someone will object to our intervention And so a host of other factors must be relevant to such a decision and, Williams argues, must not be self-righteously ignored in favor only of "what is the simply right thing to do " For a *state* will be doing whatever some group thinks ought to be done, and that means (a) that the intervention will have many repercussions for other states and it would be irresponsible not to consider those consequences *ex ante*; (b) there is, sadly, suffering of this general sort going on in more than one place at any given time and one has to have a plausible account of why this intervention and not some other (or all others) demands our attention and action, (c) in the greatest contrast to the private morality of rescue, a state decision means that some individuals are deciding that other individuals should be put at great risk, and surely the considerations relevant to that disanalogy alter what should count as reasons to intervene, and (d) modern liberal governments are subject to democratic will If its leaders are responsive to some moral consideration, they must also, as Williams puts it in his "realistic" way, "be able to get away with it " It must "play in Peoria" (151) And this is altogether proper, not the self-serving or temporizing dodge the moral point of view sometimes suggests

The most important issues in the book are raised in the middle chapters and concern the questions of what it means to consider freedom an essentially "political value," why we should view it that way, and how thinking of it that way changes our view of its possible competition with equality Williams disagrees with those like Ronald Dworkin who argue that no one has a "claim in liberty" when he is

simply being denied the capacity to do what he wants He has such a claim only when he is denied what he is entitled to do On Dworkin's view, this means that when people complain that the state is restricting their liberty by constraining them in nonentitled domains, the proper response is to explain to them that, appearances to the contrary, this is not so, that they are "really" as free as before the restriction Williams finds this all very implausible and dangerously close to the controversial "force to be free" arguments of Rousseau (He has similar objections to Quentin Skinner on republican liberty Compelling participation in public offices might be a good thing, but it should count as a loss in liberty) Political life often *does* demand a real loss of liberty for the sake of other competing goods, and we should acknowledge this and try to justify it, not try to reeducate the complainants

Williams follows the same strategy in understanding the claims of equality as "political claims " For Williams, not treating persons differentially based on skin color is not a matter of recognizing their common status as free and rational beings, or as their having moral capacities, as if these capacities are a special sort, always independent of contingent differences Not treating persons unequally (in the relevant respects, especially with regard to equality of opportunity and equality of respect) is simply not to exercise one's will arbitrarily, "like that of some Caligulan rules who decided to execute everyone whose name contained three R's" (100) We do not need to pay the "high price" (102) of appealing to "transcendental" properties in order to rule out such arbitrariness

However, in cashing out what he means by "arbitrarily" and in explaining the nature of the wrongness of so exercising our will, Williams appeals to a notion of "relevant reasons" that seems to rest on the moral considerations that he is trying to avoid The "weak" principle he begins with is that "for every difference in the way people are treated, a reason should be given," and he goes on to insist that these reasons must be relevant and they must be "socially operative" (equality before the law does not does not count as such if only wealthy people can make effective use of such an entitlement) (107) But this sort of prohibition on unequal treatment (in the relevant respects) appeals to the fact that persons are reason-giving and reason-demanding in a way that is normatively significant, and that they must be treated as such beings independently of how well or poorly they can be said to exercise such reason What is done to them or what is foreclosed to them must be in principle defensible *to* them because they enjoy a status that can only be called "moral," and that in a way that does not look in any significant sense historically inflected

or contingently variable Indeed, Williams says that with this notion of the relevant claims of equality and its grounds, we "have a genuine moral weapon" to wield against unequal treatment (107)

This though is not where the political considerations that interest Williams enter He argues that we will not have a firm grasp of *what* being committed to achieving equality of opportunity consists in unless we consider the political dimensions of trying to realize such a goal When we consider the realization issue, we have to concede that various dimensions of achieving equality of opportunity will be in conflict with, say, equality of respect Some distribution schemes will be patronizing and will damage the competitive chances of others, some schemes for insuring equality of respect, if indifferent to issues like the human desire for prestige, and the economic need for certain rare skills, will be doomed to a "futile Utopianism" (114) All we can hope for is the "best way of eating and having as much cake as possible" (114) (There is a famous German word for this "*Aufhebung*" or "sublation," and Williams indicates at several places that he appreciates the Hegelian resonances to what he is saying (25, 31, 36, 48))

This is another example of the implications of the phrase from Goethe's *Faust* that Williams takes as his watchword, "In the Beginning Was the Deed " He means by invoking it to claim that no principle can by itself determine its own application and that the issue of the "actual" content of any principle, when it is applied and asserted in the public domain, when it is opposed by others and directed towards the use of coercive restraint on others, is not a secondary or supplementary issue, does not involve a mere unfortunate compromise with an imperfect world It is *in* such contestations and in the midst of historical change that the concept can be said to have any determinate content at all (The implications of these sorts of claims for philosophy, as it is now practiced professionally, are obviously profound, something only occasionally hinted at here)

In his discussion of a possibly "left Wittgensteinianism" (Wittgenstein quoted the Faust passage in *On Certainty*), Williams notes the important point that this does *not* mean that our actual practices should be said to be "*grounded*" in something like a "description of how we go on," as if the latter could be given as a reason to someone who requested one *That* too would have to count as the way we now go on, it could not justify anything "foundationally " Nothing could (This is the paradox of "conservative Wittgensteinians" who must invoke such a theory to rail against theory, and it is why Williams is not, say, like Burke, or like the Southern segregationists who criticized moral demands for integration by appealing to the "futile Utopianism" of integration and to the priority of "how we've always

gone on " No one gets to be the unchallengeable arbiter of what is now possible, that too is always up for debate) Nor is his position that our practices are "groundless" in the sense of arbitrary, or justifiable only relativistically That would be to accept in a question-begging way the universalistic and a priorist criterion of justification and to count everything that fell short of that as unjustifiable, assertable only ironically (Williams thinks that Richard Rorty is guilty of this) Political practices are justifiable (or not) to the persons they affect in a context, at a time, in the light of other competing claims and opposing norms, given a historical location in an ongoing development In our location, we are self-conscious about all of this, and that makes for a much more complicated situation

The point Williams is making is that the social and ethical "we" is not like Wittgenstein's linguistic-community "we " The latter refers to anyone who can speak a language, with whom we might be able to converse, and in that sense there is no human who stands outside such "we " But political communities are different because so contentious, indeed our community especially can no longer count as a "satisfactorily functioning whole" (37) Some widely accepted parts of it might even stand condemned by "perfectly plausible extrapolations of other parts" (37)

This might have a MacIntyrean sound to it, can sound like a claim about the normative incoherence of "modernity " But Williams does not think that because the political community is not functioning as a whole, it is not functioning It is not an illusion that some parts stand condemned by other parts, the condemnation is true, and we then have the complex political task of bringing some order to our normative lives No one gets to decree that any such order is just not possible "any more " So, from this point of view, when Williams says something like, "Whether it is a matter of philosophical good sense to treat a certain practice as a violation of human rights, and whether it is politically good sense, cannot ultimately constitute two separate questions" (72), he is not encouraging us to trim our philosophical sails to accommodate prevailing political winds He is trying to point out that it is a very bad sailor indeed who proposes to sail to his goal no matter the wind, just as bad as one who lets the prevailing winds and currents set his destination To philosophers who profess no interest in actual sailing (just its "principles") Williams makes very clear here why he is as suspicious as Hegel of such attempts to instruct the political world from on high about how it ought to be

ROBERT B PIPPIN

University of Chicago

The Sense of the Past Essays in the History of Philosophy BERNARD WILLIAMS Edited with an introduction by MYLES BURNYEAT Princeton Princeton University Press, 2006 xxii + 393 p Cloth $39 50

Take it all in all, the book is a corker If you care for the history of philosophy, then either you have already acquired it or else you are going to

What you have got, or will get, is a collection of twenty-five essays, all but two of which have been published before They were seen through the press by Myles Burnyeat, who has added a touching introduction and a handful of notes on Greek terms Otherwise (I think) there are no changes The essays have been reset, but their original pagination is not indicated There is a list of Williams's philosophical publications, but no bibliography of works referred to The book wants an index [1]

Fourteen of the pieces deal with ancient thought Plato gets seven, Aristotle four, Sophocles one, Homer one—and the first is on "The Legacy of Greek Philosophy " There are also four papers on Nietzsche, three on Descartes, and one each on Hume, Sidgwick, Collingwood, and Wittgenstein A half-dozen of them were written by Professor Williams of Oxbridge They are introductory essays It is hard to imagine how they could have been better done Another half-dozen were composed by Saint Bernard of Sils Maria—they touch on horror, unbearable suffering, *et caetera* Some will regard them as the high points of the collection The rest—the majority—were written by Bernard Williams, that most subtle and stylish of professional philosophers, the author of (*inter alia*) *Descartes* and *Shame and Necessity*, a man who could not see a nail without striking it squarely on the head

The essays span forty years, they are spread over a variety of topics, and they are written in three registers But if they do not make a homogeneous book, they are held together by some general themes—and Williams had planned "a substantial introductory essay on the history of philosophy" (ix) The essay was never written, but its general lines are no secret,[2] and since Williams's views, in one variant or

[1] Space might be found by dropping chapter 6, virtually all of which reappears *verbatim* in chapter 8 There are other, less startling, repetitions across the essays

[2] See especially *Descartes The Project of Pure Enquiry* (New York Penguin, 1978), pp 9–11, "Descartes and the Historiography of Philosophy" (257–59), and the note by Adrian Moore, "who knew Bernard's thoughts well" (ix–x)

0022-362X/07/0410/540–45 © 2007 The Journal of Philosophy, Inc

another, are widely accepted among philosophers, it may be both useful and diverting to say something about them [3]

The views sometimes parade beneath the slogan "The history of philosophy is a part of philosophy", or as Williams might have put it "The history of philosophy is in the first instance philosophy" (ix) The slogan intimates that the relation between philosophy and its history is significantly different from the relation between arithmetic and its history, or astrology and its history—Williams speaks of "that involvement with its own history which is characteristic of philosophy and not of the sciences" (8) What he intends by that may be elaborated in three stages first, there is a fundamental distinction, then a swipe at analytical philosophy, and finally a Nietzschean thesis

First, then, Williams distinguishes between history of ideas and history of philosophy The nomenclature is cumbersome (and unilluminating) I shall speak, inelegantly, of HI and, saucily, of HP Williams once allows that HI "certainly has nothing wrong with it" (260), but his Nietzschean thesis implies that it is a Bad Thing As for HP, "very often, [it] is merely constructed out of our preconceptions" (261) that sounds bad, but he must mean that HP (like everything else in philosophy) is often badly done—HP in itself is a Good Thing

The two histories are distinguished along two axes like knowledge and opinion in Plato's *Republic* they have different products and they have different objects First, HI "yields something that is history before it is philosophy, while with the history of philosophy it is the other way round" (257) Second, HI "naturally looks sideways to the context of a philosopher's ideas," whereas HP "has an interest in diachronic influence" and "is more concerned to relate a philosopher's conception to present problems" (257) So it looks as though HP is—or includes—what Germans call *Rezeptionsgeschichte* HI wonders where a dead philosopher got his ideas from, HP wonders what happened to them after his death

There are two different things there—but why not do both? According to Williams, you cannot, for two reasons First, "the best possible history of ideas is likely to show that the philosophy did not in fact mean in contemporary terms what subsequent philosophy has most made of it" (257) For "if philosophers are going to be

[3] There is a literature on the subject, and any number of *obiter dicta* I have learned much from Kevin Mulligan, "Sur l'histoire de l'approche analytique de l'histoire de la philosophie de Bolzano et Brentano à Bennett et Barnes," in J -M Vienne, ed , *Philosophie Analytique et Histoire de la Philosophie* (Paris J Vrin, 1997), pp 61–103, and also from the introduction to Anthony Kenny's *New History of Western Philosophy, Volume I Ancient Philosophy* (New York Oxford, 2004)

influential, it is as well that they should be misunderstood" (25) It is a fine aphorism (and not the only one in the book), but it is not true—and even if it were, I cannot see why that should prevent anyone from indulging both in HI and in HP Second, "the kinds of sensibility needed for the two activities are bound to yield partly incompatible products" (257) They will yield different products But how could they possibly yield incompatible products?

It will be objected that I have misrepresented Williams despite what he says, HP has nothing to do with diachronic influence (something in which Williams himself shows no interest) Rather, HP has two products one is philosophy and the other is not *Rezeptionsgeschichte* but textual exegesis I think that is true Even so, Williams's arguments do not show that HI and HP are incompatible professions And they are not

Next, stage two Williams affirms that the two elements in HP "do not simply, or always easily, go together" (258), and to prove the point he cites the case of "analytical history of philosophy " By that he seems to mean "history of philosophy done by analytical philosophers", and the analytical philosophers seem to be those old Oxonian devils R**e and A****n and their black acolytes (Some readers may think that Williams is attacking Strawman) In any event,

> at its extreme, this activity could take the form of triumphant anachronism, as when it used to be said, in the heyday of analytic confidence, that we should approach the works of Plato as though they had appeared in last month's issue of *Mind* (258) [4]

After cataloguing a number of "charges against analytical history of philosophy," Williams concludes that "this activity has no title to being history of any sort" (258)

The charges are as silly as they are familiar, but perhaps the concluding verdict is just? When I was a lad, my first philosophy tutor—Dick Hare—did indeed tell me to read Plato as though he had just been published by the Clarendon Press I was supposed to be learning philosophy It was Hare's idea—perhaps not wildly eccentric—that I might profitably read what some philosophers had written My reading list—the subject was the relation between knowledge and belief—included the *Meno* and something by A J Ayer, as well as a few items by lesser stars I was not meant to worry too much about the opinions of Freddie Ayer I was doing philosophy, not history of contemporary philosophy Ditto, *mutatis mutandis*, for Plato The

[4] Elsewhere, the advice is called "Ryle's frequent injunction" (344)

essay I wrote was not a piece of (jejune) history of philosophy It was a piece of (extraordinarily jejune) philosophy

Things were similar for the grown-ups Williams says that Strawson said that Grice said that "all Oxford philosophers would agree that we 'should treat great but dead philosophers as we treat great and living philosophers, as having something to say to *us*'" (344) That is how Strawson himself treated Leibniz in his *Individuals*

You might agree with Williams that such an activity has no title to being called history It scarcely matters—and so far as I can recall, the analytical devils never did call it history (There was something else they called history of philosophy—I shall come back to that)

You might also think that the activity is a peculiar one Why read the works of a defunct foreigner when what you are after is a clutch of eternal truths? But Williams does not find the activity peculiar in that respect For—this is now stage three, the Nietzschean thesis—he urges that the point of HP is to be found in "the possibility of the past philosophy's being untimely, and helping to make strange what is familiar in our own assumptions" (263) Or again "The point of reading philosophers of the past is to find in them something different from the present" (344) And more expansively

> To justify its existence, [HP] must maintain a historical distance from the present, and it must do this in terms that sustain its identity as philosophy It is just to this extent that it can indeed be useful, because it is just to this extent that it can help us to deploy ideas of the past in order to understand our own (259)

I call this a Nietzschean thesis because Williams offers it as a special case of a general theorem, which he ascribes to Nietzsche "the point of any history is to achieve some distance from the present, which can help one to understand the present" (258)

If that is the real HP, how does it differ from the degenerate activity of the analytical philosophers? Adrian Moore says that the point of HP

> was not, as philosophers in the analytic tradition used to think, to indicate voices of yore which could be heard as participating in contemporary debates precisely not It was to indicate voices of yore which could *not* be heard as participating in contemporary debates, and which thereby called into question whatever assumptions made contemporary debates possible (ix)

That is a neat antithesis, but a false one In truth, the devils and the angels agree that the great dead can (and perhaps should) participate in contemporary debates Their difference is rather this in answer to the question "Why read the *Theaetetus*?" the devils say "It

may help you to get clearer about the nature of knowledge," and the angels assert "It may help you to get clearer about the nature of knowledge, because it is so distant from anything familiar to you." Demonic HP may call on any historical text which bears on a contemporary issue. Angelic HP will call on an historical text only when it is pertinent to a contemporary issue and also offers a point of view which is radically different from any contemporary point of view. So angelic HP is a part of demonic HP, and if the latter is not history, then neither is the former.

However that may be, HP—as Williams characterizes it—is unalluring. After all, if you are an occidental philosopher and you long for something rich and strange, why on earth go to ancient Greece? Why not go to China? And in any case, the Nietzschean notion which Williams invokes is grotesque. There are several reasons for studying the past. One of them—common enough, I should have thought, and decent enough—is to learn about the past. Of course, you may not want to learn about the past—but in that case you are not an historian.

Fortunately, Williams does not practice the HP he preaches. As he says, in *Shame and Necessity* he "concentrated on those ideas about which I was convinced that they bear a rather close relation to our own" (64). Moreover, when he does sniff out something strange, he does not bark and wag his tail. An example. Aristotle allegedly makes a close connection between ethics and biology, according to Williams, "we do not share this picture, and it is a denial of history and of scientific change to pretend that we do" (197). "Aha," you say, "just the sort of thing he's after. no doubt he'll sink his teeth into it." Not at all. he looks down his nose and trots away.

Then what was Williams up to? David Charles says of an activity which he calls "philosophical scholarship" that it

> aims to represent Aristotle's discussion as focused, where appropriate, on questions which also interest contemporary theorists, and to assess the philosophical significance of his answers by comparing them with solutions put forward today.[5]

Suppress the insinuation that the philosophical assessment of a dead author is only done by comparison with his modern successors, and you have a plausible account of what Williams is usually trying to do. Drop the restriction to questions which interest contemporary theorists, and you arrive at something which describes well enough what most anglophone historians of philosophy do most of the time.

[5] Charles, *Aristotle's Philosophy of Action* (Ithaca, NY: Cornell, 1984), p. ix.

You might call it philosophical exegesis Has it a claim to the title of history? Just about, I would say—but who cares? History or not, it the sort of thing I do myself, it is more or less what the Oxonian analyticals called history of philosophy, and it is what Williams himself does—and does with consummate artistry

What's the point of it? Anthony Kenny says that "we may read the philosophers of other ages to help to resolve philosophical problems of abiding concern, or to enter more fully into the intellectual world of a bygone era" (*op cit*, p xi, the disjunction is inclusive) Perhaps there is also a third reason for reading the great dead Recently I spent some time working on Frege's *Begriffsschrift* Did I hope to "resolve philosophical problems of abiding concern"? No I read Ian Rumfitt for that sort of thing Did I want to "enter more fully" into a bygone world? (Return with us now to those thrilling days of yesteryear) No for that I read Lothar Kreiser So what did I want? Well, I wanted to understand what Frege had said

JONATHAN BARNES

Ceaulmont

Philosophy as a Humanistic Discipline BERNARD WILLIAMS Selected, edited, and with an introduction by A W MOORE Princeton Princeton University Press, 2006 xx + 227 p Cloth $35 00

That this collection includes essays which span Bernard Williams's career, dating back to his first published essay, "Tertullian's Paradox" in 1955, may obscure what is most compelling about them that the essays, which explicitly reflect on philosophy were written during a period in which Williams knew he was dying I do not intend this remark to be sentimental He thought philosophy (the profession) was moving in a bad direction (away from philosophy), and he devoted a significant chunk of the time he had left trying to reorient it As befits the man, he was charmingly skeptical of his own activity He quotes—I can hear his voice—"a marvelous passage" from Nietzsche "It quite often happens that the old man is subject to the delusion of a great moral renewal and rebirth, and from his experience he passes judgments on the work and course of his life, as if he had only now become clear-sighted, and yet the inspiration behind this feeling of well being and these confident judgments is not wisdom, but *weariness*" (198)

Weariness Boredom People who knew Bernard regularly said they were afraid of boring him My own sense is that they mistook for boredom what was really impatience—(as expressed by a brilliant Oxford-educated Englishman who came to maturity in the middle of the last century) Boredom is one of the zero-settings of human engagement when nothing seems any more worth doing than anything else Impatience emerges from the increasingly urgent sense that there are things worth doing *rather than* the present conversation This is the experience that made his interlocutors anxious And it is philosophically significant that it should have been impatience, not boredom Bernard Williams was an atheist, and he tried to work out—in thought and in life—what this must mean He correctly diagnosed the covert religiosity in the claim, 'If God does not exist, everything is permitted' as though God was required for any legitimate sense of constraint The idea of God, for Williams, is a thoroughly human invention—and yet, for him, some things should be permitted, others forbidden, others encouraged, others frowned upon, others denounced To understand how these distinctions could legitimately be made was, Williams thought, a proper task of

0022-362X/07/0410/546–50 © 2007 The Journal of Philosophy, Inc

philosophy Above all, he thought we had to abandon any vestige of the idea that the cosmos is cheering us on (or looking down on us in disapproval) The cosmos does not have a point of view, and the idea that there is any absolute measure of our importance in the universe is illusion This illusion is present even in the view that our cosmic importance is vanishingly small Williams diagnoses this as a muddled form of sentimentality, and he again turns to his modern philosophical hero "Nietzsche by contrast got it right when he said that once upon a time there was a star in a corner of the universe, and a planet circling that star, and on it some clever creatures who invented knowledge, and then they died, and the star went out, and it was as though nothing had happened" (138) Though thoroughly anti-religious, there is a rigor in Williams's thought that reminds me of certain religious writers Just as some discipline their entire lives around the idea of God, Williams's entire intellectual output is organized around the idea that whatever significance is to be found in the human endeavor, it must be found as somehow emerging from that endeavor Williams's sense of urgency, which grounded his impatience, flowed from a solidarity with that endeavor He was committed to the idea that though truthful inquiry would expose much that was vain, pompous, and wishfully deluded in human life, there would nevertheless emerge genuine answers to the question, 'How should one live?' There was no time for bullshit

Philosophy, for Williams, is a humanistic discipline in the sense that it is "part of a more general attempt to make the best sense of our life, and so of our intellectual activities, in the situation in which we find ourselves" (182) Much as Plato tried to characterize philosophy via a constitutive contrast with sophistry, the contrast of the current era, Williams thinks, is with scientism Scientism is a name he gave to a family of misuses of the image and authority of science Williams had the highest regard for science and he thought that philosophy should inquire into it There need be no firm line between, say, physics, logic, linguistics, and the philosophical inquiries into them, thus there was every reason for some philosophical papers to have a scientific style "Scientism is, rather, a misunderstanding of the relations between philosophy and the natural sciences which tends to assimilate philosophy to the aims, or at least the manners of the sciences" (182) Williams was a naturalist, in the sense that we ought to understand humans as thoroughly part of the natural world, but he also thought that a false image of naturalism constrained and distorted our sense of what inquiry into the human should look like And he wondered whether scientism contributes to "the well-known and highly typical style of many texts in analytic philosophy which

seeks precision by total mind control" which seeks "to remove in advance every conceivable misunderstanding or misinterpretation or objection, including those that would occur only to the malicious or the clinically literal-minded This activity is often rather mournfully equated with the boasted clarity and rigor of analytic philosophy" (183) There were, he thought, various currents in the contemporary profession which tended "to implant in philosophers an intimidatingly nit-picking superego, a blend of their most impressive teachers and their most competitive colleagues, which guides their writing by means of constant anticipation of guilt and shame" (183)

But the problem with scientism goes beyond style Williams is particularly concerned with a tendency to overlook the crucial role of history for philosophy The history of science tends to be *vindicatory* in the sense that a later theory tends to make sense of itself and its predecessors in terms of which holders of the later as well as earlier theories have reason to view as an improvement So, while the history of science is of immense intellectual interest, there is no reason to see it as constitutive of contemporary scientific inquiry Science, thus, can detach itself from its history No such analogous move can properly be made by philosophy The concepts with which we understand the human condition—anthropological, ethical, political—have histories that are not vindicatory in this sense There are at least three consequences of philosophical interest First, the attempt to derive a concept from timeless human needs or from the a priori conditions of a finite rational being will be at best partial, at worst misleading For, crucial aspects of the concept will have arisen out of contingent historical circumstances that escape any such derivation An example Williams gives is the virtue of truthfulness while there may be a universal human need for communications that are sincere and accurate, the forms of integrity and authenticity that our conception of truthfulness has are the result of contingent historical development To understand what truthfulness is we need to understand how it became what it has become Second, to understand the political and ethical concepts with which we understand ourselves, we have to see them as having won out over competitors—but not by having won an argument that both sides could recognize in vindicatory fashion The liberal justification of liberalism, again to take Williams's example, is not a form of argument that will appear justificatory from outside a liberal perspective (Ask M Charlus about the equal rights to life, liberty, and the pursuit of happiness, and get ready for a hilarious diatribe) Liberalism "won," Williams thinks, due to a series of historical crises, the outcomes of which were contingent This recognition need not be reflectively destabilizing, Williams

thinks, but if we are to understand the proper reflective attitude to take towards our own conceptions, we need to know how they came to occupy the place they have It is another illusion of scientism, he argues, to think that what we ought to be doing is providing timeless justifications from an absolute point of view of our ethical and political ideas There is no such justification, but it is only from a scientistic perspective that that is experienced as a disappointment—as though that recognition forces us to accept something second-best Third, it is a feature of modern political life that its most basic ideas are presented as somehow timeless, without a future beyond themselves Williams considers the idea that every person is due equal consideration On the one hand, we recognize that most people in the past have not believed this and that many people today do not, on the other hand, within the modern liberal perspective, it is inconceivable that there should be a viable political future without it A sense of timelessness is a manifestation that we are at the outer bounds of our political outlook, of what we could possibly see beyond Williams does not think it a criticism of liberalism that it cannot see beyond the limits of its most basic convictions But its indifference to why that is so—why what is most basic seems to be just there—betrays philosophical complacency, a lack of interest in historical forms of reflective self-understanding

Williams is famously associated with the philosophical virtue of, as he put it, "getting it right" the painstaking, often meticulous, work involved in distinguishing truth from falsity This he clearly thought worth doing—even though the universe did not give us this (or any other) task And he, as much as anyone I have known, admired and tried to live up to the ideals of clarity and precision embodied in analytic philosophy But he contested the idea that there was one style or approach that set the standard for precision, clarity, getting it right And he, in effect, reintroduced a concern with the *how* of philosophy Philosophy not only needs to get it right, it needs to *ring true* and this is essentially an imaginative accomplishment "It would be a bad friend of philosophy who did not admit that there is quite a lot of philosophical work that is unrewarding by any standard unhelpful, boring, sterile The awful fact is that some of hardly tries to be anything else" (204) If we think of some of the greatest works in the philosophical tradition—by Plato, Descartes, Kant, Kierkegaard, Hegel, and Wittgenstein, for example—the form of presentation is not just a clever literary device for entertaining the reader, but is essential to the philosophical activity itself It has been a historical contingency of recent times—which we have not noticed, given that it is also contingently a period in which we have tended not to notice

historical contingencies—that we have ignored the philosophically crucial issue of the form of communication

Williams admired Nietzsche more than I do, and he had an indefatigable impatience for Aristotle, a philosopher whom I hold in awe Williams is evidently so delighted by Nietzsche's sensibility—a companion spirit in a universe rather thinly populated with such spirits—that he occasionally attributes to him powers that I find wishful So, for instance, he says that Nietzsche "mined or booby-trapped his works against being assimilated" to dominant fashions (208) A herd of Nietzscheans is indeed an unlovely sight, but if there have been any bombs going off under their hooves, I have not noticed It would seem to me more accurate to say that Nietzsche gave us ironic *clues* as to how he is to be read He did make it possible for someone with the brilliance and sensibility of Bernard Williams to savor him As for Aristotle, it would irritate Williams no end if I teased him by telling him that there is something about his style that reminds me of the Stagirite But, *for Williams,* getting it right often consists in warding off excesses on either side of what he takes to be true So, to close with vintage Williams in defending the idea of an absolute conception of reality he says, "there are two mistakes to hand here One is to suppose that just because there is an uncontentious sense in which all our conceptions are ours, it simply follows from this that they are all equally local or perspectival, and that no contrast in this respect could conceivably be drawn from inside our thought between, for instance, the concepts of physics and the concepts of politics or ethics The other mistake is to suppose that if there is such a contrast, and one set of these concepts, those of physical science, are potentially universal in their uptake and usefulness, then it follows from this that they are somehow intrinsically superior to the more local conceptions which are humanly and perhaps historically grounded The latter is scientistic error " (187) This looks an awful lot like the virtue of truthfulness, magnificently displayed throughout Williams's work by striking the mean If one wants to think anew about philosophy and its possibilities this book is required reading

JONATHAN LEAR

University of Chicago

CAMBRIDGE

New and Noteworthy Titles

The Metaphysics of Everyday Life
An Essay in Practical Realism
Lynne Rudder Baker
Cambridge Studies in Philosophy
$85 00 Hb 978 0-521-88049-7 272 pp

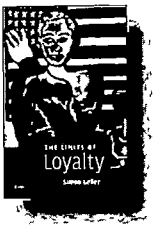
Patriarchal Religion, Sexuality, and Gender
A Critique of New Natural Law
Nicholas Bamforth and David A J Richards
$95 00 Hb 978-0-521 86863-1 416 pp

Music, Philosophy, and Modernity
Andrew Bowie
Modern European Philosophy
$99 00 Hb 978-0-521-87734 3 432 pp

Aristotle and Beyond
Essays on Metaphysics and Ethics
Sarah Broadie
$85 00 Hb 978 0-521-87024 5 214 pp

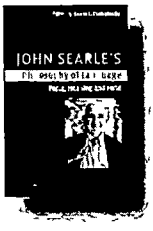
A Metaphysics for Scientific Realism
Knowing the Unobservable
Anjan Chakravartty
$85 00 Hb 978 0-521-87649-0 272 pp

The Cambridge Companion to Renaissance Philosophy
Edited by James Hankins
Cambridge Companions to Philosophy
$95 00 Hb 978 0-521-84648 6 448 pp
$29 99 Pb 978-0-521 60893 0

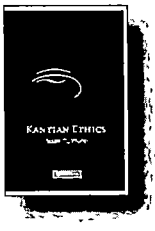


The Limits of Loyalty
Simon Keller
$85 00 Hb 978-0 521-87461-8 246 pp

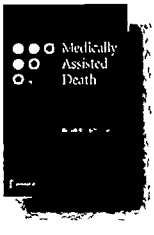
Legal Ethics and Human Dignity
David Luban
Cambridge Studies in Philosophy and Law
$90 00 Hb 978-0-521-86285-1 250 pp



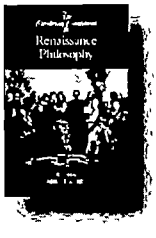
The Structural Evolution of Morality
J McKenzie Alexander
$95 00 Hb 978-0-521-87032-0 312 pp



John Searle's Philosophy of Language
Force, Meaning and Mind
Edited by Savas L Tsohatzidis
$85 00 Hb 978-0-521-86627-9 312 pp
$34 99 Pb 978-0 521-68534-4



Kantian Ethics
Allen W Wood
$80 00 Hb 978 0-521-85494-8 360 pp
$25 99 Pb 978-0-521-67114-9

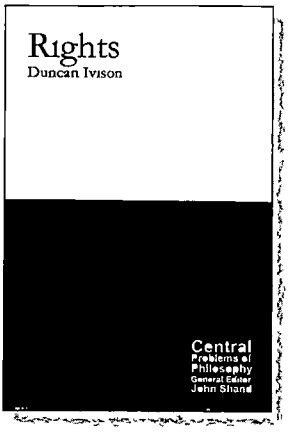


Medically Assisted Death
Robert Young
$90 00 Hb 978-0-521-88024-4 260 pp
$29 99 Pb 978 0 521-70616-2

Prices subject to change

www cambridge org/us/philosophy

CAMBRIDGE UNIVERSITY PRESS

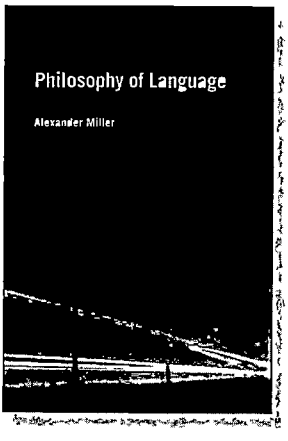


Rights

Duncan Ivison

978-0 7735-3329-5 $22 95 paper | 978-0-7735-3328-8 $80 cloth

An accessible introduction to the history and philosophy of rights

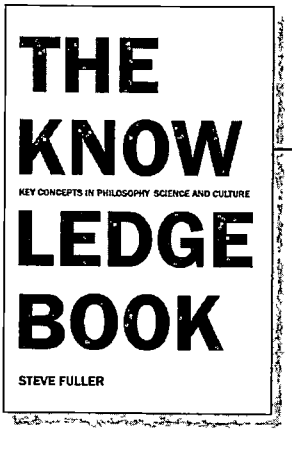


Philosophy of Language
Second Edition

Alex Miller

978-0-7735-3338-7 $22 95 paper

This second edition includes new material on Chomsky, Wittgenstein, and Davidson as well as new chapters on the causal theory of reference, possible worlds semantics, and semantic externalism



The Knowledge Book

Key Concepts in Philosophy, Science, and Culture

Steve Fuller

978-0-7735-3347-9 $22 95 paper | 978-0-7735-3346-2 $80 cloth

"An invaluable resource for anyone interested in exploring the social life of knowledge Fuller's unrivalled concision, integrated with the book's unique structure, invites readers to imagine, and to chart, the future of knowledge "

–James H Collier, Virginia Tech

MQ UP McGILL-QUEEN'S UNIVERSITY PRESS
www mqup ca

New from OXFORD

Does Feminism Discriminate Against Men?

A Debate
WARREN FARRELL and JAMES P STERBA

"The relationship between men and feminism has never been addressed in such an exciting and accessible way, or with such extensive and engaging evidence and anecdotes"—*Tom Digby*, Springfield College

(Point/Counterpoint)

2007 272 pp, 24 HT line illus cloth $39 00 paper $16 95

New in Paperback!

Symmetry and the Monster

The Story of One of the Greatest Quests of Mathematics
MARK RONAN

"Ronan does a good job of describing the mathematics in broad strokes and giving a flavor of what is happening and — more importantly — why mathematicians get excited about these questions"
—*The Mathematical Association of America*

2006 (paper 2007) 272 pp 12 halftones & line illus paper $19 95 cloth $27 00

Taking Action, Saving Lives

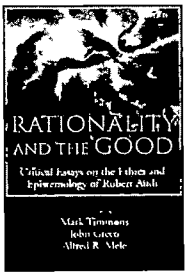
Our Duties to Protect Environmental and Public Health
KRISTIN SHRADER-FRECHETTE

In this accessible book Kristin Shrader-Frechette reveals how politicians, campaign contributors, and lobbyists have conspired to cover up environmental disease and death However, she argues that everyone, especially in a democracy, has a duty to help prevent avoidable environmental deaths

(Environmental Ethics and Science Policy Series)

2007 320 pp $29 95

Rationality and the Good

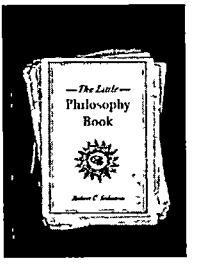


Critical Essays on the Ethics and Epistemology of Robert Audi
Edited by MARK TIMMONS, JOHN GRECO, and ALFRED R MELE

For over thirty years, Robert Audi has produced important work in ethics, epistemology, and the theory of action This volume features thirteen new critical essays on Audi by a distinguished group of authors including Roger Crisp, Walter Sinnott-Armstrong, Bernard Gert, and Raimo Tuomela

2007 288 pp cloth $99 00 paper $35 00

The Little Philosophy Book



ROBERT C SOLOMON

"Robert Solomon regularly gets directly to the heart of the matter, providing clear, unpretentious expositions and thoughtful, relevant critiques He can make the most difficult issues seem simple and straightforward, sprinkling his commentary with examples that are as timely as they are germane"—*Edward L Schoen*, Western Kentucky University

2007 112 pp cloth $30 00 paper $15 95

OXFORD UNIVERSITY PRESS

Prices are subject to change and apply only in the US
To order, or for more information, please call 1-800-451-7556
In Canada, call 1-800-387-8020 Visit our website at www oup com/us

New books from The MIT Press

Things and Places

How the Mind Connects with the World

Zenon W Pylyshyn

"This is one of the most provocative and challenging books to have appeared in philosophy or psychology in the last twenty years "
— John Campbell, University of California, Berkeley

280 pp , 26 illus $34 cloth

Time and Realism

Metaphysical and Antimetaphysical Perspectives

Yuval Dolev

"The great achievement of *Time and Realism* is that it shows how the two conflicting ideas of time misrepresent the role that temporality plays in our lives as well as in science This is contemporary philosophy at its best "
— Hilary Putnam, Harvard University

264 pp , 3 illus $30 paper

Ontology of Consciousness

Percipient Action

edited by Helmut Wautischer
foreword by Robert A F Thurman

"An essential source book for the study of consciousness and foundations of experience "
— Douglass Price-Williams, University of California, Los Angeles

672 pp , 73 illus $38 paper

Distributed Cognition and the Will

Individual Volition and Social Context

edited by Don Ross, David Spurrett, Harold Kincaid, and G Lynn Stephens

Philosophers and behavioral scientists discuss what, if anything, of the traditional concept of individual conscious will can survive recent scientific discoveries that human decision-making is distributed across different brain processes and through the social environment

368 pp , 22 illus $34 paper

Evolutionary Psychology as Maladapted Psychology

Robert C Richardson

A philosopher subjects the claims of evolutionary psychology to the evidential and methodological requirements of evolutionary biology, concluding that evolutionary psychology's explanations amount to speculation disguised as results

248 pp , 13 illus $34 cloth

Integrating Evolution and Development

From Theory to Practice

edited by Roger Sansom and Robert Brandon

Scholars argue for the importance of the developmental synthesis, or evo-devo, discussing the history and potential of this growing field of study and presenting specific case studies

336 pp , 45 illus $34 paper

Situating Semantics

Essays on the Philosophy of John Perry

edited by Michael O'Rourke and Corey Washington

Original commentary on the work of philosopher John Perry by prominent contemporary analytic philosophers, with Perry's detailed and original responses

608 pp , 6 illus $40 paper

The Letters of George Santayana

Book Eight, 1948-1952

edited and with an introduction by William G Holzberger

Letters from the last years of Santayana's life, written as he completed *Dominations and Powers*, the final volume of his autobiography, and the one-volume abridgement of his early five-part masterwork, *The Life of Reason*

704 pp $70 cloth

To order call **800-405-1619**

http://mitpress.mit.edu

THE JOURNAL OF PHILOSOPHY

VOLUME CIV, NUMBER 11
NOVEMBER 2007

page

551 *Causes That Make a Difference* — C Kenneth Waters

REVIEW ESSAYS

580 *Predication and Truth* — Tyler Burge

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J F WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Wolfgang Mann, Mary Mothersill, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C Varzi *Consulting Editors* David Albert, James T Higginbotham, Robert May, Charles D Parsons, Wilfried Sieg *Managing Editor* John Smylie

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President*, Arthur C Danto, *Vice President*, Akeel Bilgrami, *Secretary*, Michael J Mooney, *Treasurer*, Liaquat Ahamed, *Other Trustees* Lee Bollinger, Leigh S Cauman, Barbara Gimbel, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

You may also visit our website at www journalofphilosophy org

THE JOURNAL OF PHILOSOPHY

2008

SUBSCRIPTIONS (12 issues)

Individuals	$45 00
Libraries and Institutions	$100 00
Libraries and Institutions (print and online)	$180 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for advertising rates, ad space is limited, so ad reservations are required

Published monthly as of January 1977, typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027 (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U S copyright law All rights reserved For more information about our copyright policies, please consult our January issue or contact us

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc , MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027 Periodicals postage paid at New York, NY, and other mailing offices

POSTMASTER. Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027

© Copyright 2007 by the Journal of Philosophy, Inc

ISSN 0022–362X

CENTRAL LIBRARY
UNIVERSITY OF CALCUTTA
KOLKATA CAMPUS

THE JOURNAL OF PHILOSOPHY
VOLUME CIV, NO 11, NOVEMBER 2007

CAUSES THAT MAKE A DIFFERENCE*

The best theories of causation do a better job clarifying what it means to identify a cause than specifying what it means to pick out the cause that made the difference For instance, counterfactual theories can explain what it means to say "Mary's striking the match caused the match to light," but they do not explain what it could mean to specify "Mary's striking the match, rather than the presence of oxygen, caused the match to light " Mary's striking the match counts as a cause on counterfactual accounts because if Mary had not struck the match, then the match would not have lit But the same reasoning leads to the conclusion that the presence of oxygen was also a cause, if oxygen had not been present, then the match would not have lit [1] There is no denying that oxygen was a real cause Nevertheless, in ordinary contexts, observers would identify Mary's striking the match, and not the presence of oxygen, as the factor that made the difference And in an important sense they would be right The puzzle is, in what sense?

* Thanks to Jeff Barrett, Mark Borrello, Richard Creath, Norman Dahl, Richard Grandy, Bill Hanson, Susan Hawthorne, Toben Lafrancois, Roberta Millstein, Bart Moffat, Jane Maienschein, Sandy Mitchell, Katie Plaisance, Chuck Stieg, Marcel Weber, Beyong-Uk Yi, and especially to Helen Longino, Jay Odenbaugh, Michael Strevens, Jim Woodward, and referees of this JOURNAL for constructive comments on various versions of this paper

[1] Some might say that Mary's striking the match is the "triggering cause" because it explains the timing of the event and that other causes, including the presence of oxygen, are "structural " But this distinction breaks down under a counterfactual analysis of causation For example, under a counterfactual analysis, the timing of the match lighting was also due to the fact that oxygen was not removed just before the match was struck and then reintroduced with a torch just after the match was struck In this case, the match would have lit at a later time Hence, one might argue that the timing of the match lighting was caused by such a series of events not occurring The distinction between triggering and structural causes, or something much like it, can be sustained by analyzing it in terms of actual difference making causes, rather than mere causes

0022-362X/07/0411/551–79 © 2007 The Journal of Philosophy, Inc

There is a distinction to be drawn between identifying the causes of an actual phenomenon, such as the presence of oxygen, and picking out the cause that actually made the difference, such as Mary's striking the match This distinction is elusive, and it is tempting to dismiss it as illusory On this view, there is an ontological difference between causes and noncauses, but there is no ontological difference among causes This was the position taken by Mill who argued that the capricious manner in which we select causes from conditions indicates that there is no scientific ground for a distinction [2] Subsequent attempts to explain the selection of "the cause" from many causes (or from many "conditions") have generally centered on interest-laden factors that depend on the context of inquiry For example, H L.A. Hart and A M Honoré say that the selection depends on principles that vary "both with the type of causal question at issue and the circumstances in which causal questions arise "[3] As J L Mackie explained

> since even the choice of a field is relative to a purpose or a point of view, and since even apart from this what we recognize as a cause, rather than a mere condition, commonly depends on what we know—or what we knew first—or what is closely related to our interests, there is much to be said for Mill's refusal to distinguish 'philosophically speaking' between causes and conditions As an analysis of ordinary language, this would be wrong, but from a theoretical point of view, as an account of causal processes themselves, it would be right [4]

Mackie's claim has not been seriously challenged According to the received philosophical wisdom, the causes we select as "the cause" are the causes that are most relevant to whatever interests we happen to have Selected causes do not share some ontological feature that nonselected causes lack What selected causes share is our interest And our interests, as Hart and Honoré's analysis of the contextual factors underlying the selection of causes reveals, vary from one epistemic context to the next Ontologically speaking, causes are causes This view suggests that seeking the one cause among many that actually made the difference would be a fool's errand Although this is the dominant view in the philosophical literature, it is difficult to sustain when attention is turned to scientists' concrete efforts to identify which cause among a set of causes is actually making the difference

Nowhere is the practice of picking out actual difference makers more real than in biologists' account of the role of DNA in the devel-

[2] Mill, *A System of Logic* (New York Macmillan, 1950), p 244
[3] Hart and Honoré, *Causation in the Law* (New York Oxford, 1985), p 31
[4] Mackie, *The Cement of the Universe* (New York Oxford, 1980), p 120

opment and functioning of cells and organisms Biologists agree that DNA exercises its roles through the production of RNA and polypeptide molecules Given the mechanisms of RNA and polypeptide syntheses, DNA is obviously a cause of linear sequences of nucleotides and amino acids comprising these molecules But, just as obviously, DNA is only one of many causes Although biologists and philosophers often talk as if DNA "produces" RNA and polypeptides, everyone knows that many different molecules and cellular structures play necessary roles in the *in vivo* syntheses of these important molecules Even in the simplest case, the production of RNA molecules in prokaryotes, the syntheses causally depend on RNA polymerase This has led some philosophers to argue for "parity" and against the "reductionistic bias" of biologists who center so much attention on the role of genes and DNA [5] But perhaps biologists are right, perhaps there is an ontological feature that distinguishes DNA as the actual difference maker Determining whether biologists are right requires solving the conceptual puzzle about the difference between what it means to be a cause and what it means to be a cause that actually makes a difference The aim of this paper is to solve this conceptual puzzle and address two related questions about causal reasoning in biology

The first question concerns arguments for causal parity in complex biological systems These arguments typically start with the premise that the kind of element emphasized by scientists as the cause of a given process is actually *just* one of many causes of the process Parity arguments then claim that picking out one cause, when in fact there are many, cannot be justified on ontological grounds because, after all, causes are causes Some proponents of parity acknowledge that picking out one cause among many might be a useful short-term heuristic [6] Others argue that it represents an unjustifiable bias that "privileges" false reductive accounts of complex processes over true holistic ones [7] Either way, if the basic pattern of reasoning in parity arguments is sound, then picking out DNA as the actual difference making cause of the linear sequences of nucleotides in RNA and amino acids in polypeptides cannot be justified on ontological grounds Biologists, however, have been quick to identify DNA (in most con-

[5] For a number of essays and references to essays arguing for "parity," see Susan Oyama, Paul E Griffiths, and Russell D Gray, eds , *Cycles of Contingency* (Cambridge MIT, 2001)

[6] For example, see Oyama, *Evolution's Eye* (Durham, NC Duke, 2000)

[7] For example, see Lenny Moss, *What Genes Can't Do* (Cambridge MIT, 2003), and Jason Scott Roberts, *Embryology, Epigenesis, and Evolution Taking Development Seriously* (New York Cambridge, 2004)

texts) as the cause of these linear sequences My first question about causal reasoning in biology is Does the existence of multiple causes in the complex processes of biological development imply that there cannot be an ontological difference that justifies focusing attention on genes and DNA?

The second question concerns how causal generalizations in biological sciences, such as molecular biology, differ from causal generalizations in many physical sciences In physical sciences, investigators often prize causal generalizations that hold over as wide a range of conditions as possible Indeed, many philosophers of science claim that physical scientists seek causal generalizations which hold over all possible conditions There is an active debate concerning whether such universality is an appropriate ideal for physical sciences, but probably everyone, or at least nearly everyone, would concede that physical scientists are seldom satisfied with generalizations that do not hold under a wide variety of artificial and natural conditions and continuously seek generalizations with ever broader ranges of application Biologists are also concerned about the range of conditions under which their causal generalizations hold But in contrast to physicists, they seem *perfectly* satisfied with causal generalizations that are highly sensitive (that is, those which do not hold over a wide range of conditions) Why is the ideal of universality so much less important to biologists than to physical scientists? Have biologists set their sights too low as some philosophers contend? And if these philosophers are wrong (as almost every philosopher of biology claims), that is, if universality is not an appropriate ideal for biological sciences, then what ideal is? My second question about causal reasoning in biology is What kind of causal generality matters in the context of biology?

I will solve the conceptual puzzle about actual difference making and address the questions about causal reasoning in biology by analyzing causal concepts that have not been adequately analyzed, namely, the concepts of *potential difference maker, the actual difference maker,* and *an actual difference maker* These concepts, and my explications of them, are derivative in the sense that they presuppose an underlying notion of cause My explications are based on James Woodward's manipulability theory of causation, which provides a useful philosophical framework for getting clear about causal reasoning in science I will not address the question of whether my analysis could be preserved under alternatives to this theory I contend, however, that if a theory of causation cannot preserve my analysis of the difference between actual versus potential difference making, then it cannot be adequate for understanding causal reasoning in biology

It is worth emphasizing that I will not claim that the selection of causes is always based on identifying actual difference makers As I have said, Hart and Honoré's analysis reveals factors behind the selection of causes that vary from one epistemic context to the next But I will argue against the notion that all causes are on an ontological par Some causes are actual difference makers while others are not (regardless of epistemic context), and solving the puzzle of what it is to be an actual difference maker provides a basis for shedding new light on causal reasoning in the sciences I believe it can also shed light on the selection of causes more generally, but that is a topic for another occasion

This paper is organized as follows In section I, I examine causal reasoning in classical genetics to substantiate my claim that there is a genuine distinction between identifying a mere cause (a potential difference maker) and identifying the cause that makes the actual difference (the actual difference maker) In sections II and III, I review the basics of Woodward's manipulability theory and explain why his theory does not solve the puzzle about the difference between potential and actual difference makers I formulate the concept of *the* actual difference maker in section IV This concept, which is missing from Woodward's account, explains why it made sense for classical geneticists (and continues to make sense for contemporary geneticists) to identify one gene as "the cause" of a particular phenotypic effect even though it was already known that many genes are involved in bringing about such an effect In section V, I explain how this analysis resolves the puzzle about what it means to identify one cause among many as the actual difference maker I address the question about parity arguments, which requires specifying the concept of *an* actual difference maker as well invoking ideas about causal specificity in sections VI and VII Finally, in section VIII, I use my account of the ontology of difference makers to specify what kind of causal generality is relevant in biology

I GENES AS CAUSES IN CLASSICAL GENETICS

Classical geneticists routinely identified one gene as *the* cause of a particular phenotype while explicitly acknowledging that any phenotype was brought about by many genes, not just one Hence, this science is a good place to begin investigating how it makes sense to identify one cause among many as the actual difference maker The philosophical literature contains many inflated claims about the causal powers classical geneticists allegedly attributed to genes It is sometimes said, for instance, that Morgan and his collaborators were preformationists and believed that genes were active causal agents that individually

brought about distinctive phenotypes [8] One might try to substantiate such interpretations by searching for highly theoretical passages in which geneticists speculated on epigenesis or preformationism [9] But classical genetics was largely an experimental science, and we gain a better understanding of its basic causal reasoning by examining how geneticists explained their experimental results, not how they speculated about phenomena they could not actually explain

Experimentation in classical genetics involved carefully orchestrated breeding regimens that resulted in distinctive inheritance patterns Reasoning about these experiments invoked a conceptual division between the internal genetic makeup of an organism, called its *genotype*, and its outward character, called its *phenotype* What classical geneticists explained were the inheritance patterns formed by the transmission of phenotypic differences from one generation to another Geneticists explained these patterns by following the transmission of genotypic differences from generation to generation and attributing the presence of alternative phenotypic traits to the presence of alternative genotypes, that is, to the presence of alternative forms of genes

The classical mode of explanatory reasoning can be illustrated by examining a fragment of experimentation involving Thomas Morgan's favored model organism, the fruit fly *Drosophila melanogaster* It is worth emphasizing that the mode of reasoning illustrated by this historical example is still an important mode of reasoning in genetics today The experiment I examine below, reported in 1919, entailed breeding flies for several successive generations to produce distinctive inheritance patterns involving several different phenotypic traits, each of which was associated with a different gene The basic aim of the experiment was to investigate the precise locations of the underlying genes with respect to one another But it is not necessary for my purposes here to describe the investigative reasoning or the intricate explanation of the complex inheritance patterns involving the transmission of the several traits over half a dozen generations It suffices to examine only the fragment of the explanation involving the transmission of one of the traits over a single generation I will examine how Morgan explained the transmission of eye color from one of the experimental generations to the next

[8] See Moss, *What Genes Can't Do*

[9] Alternatively, one might perform a linguistic analysis of the terms (such as 'gene action') or a comparative analysis of metaphors (for example, see Evelyn Fox Keller, *Century of the Gene* (Cambridge Harvard, 2000)) While such approaches identify biases of classical geneticists, they do not necessarily identify the concepts doing the work in explanations of concrete phenomena My aim here is to reconstruct the reasoning that made the experimental science work in the laboratory and might reasonably be extended beyond the laboratory

The parental generation consisted of females with red eyes, the so-called "wild-type" character, and males with purple eyes, the mutant character When these flies were crossed, all offspring had red eyes So the resulting inheritance pattern was that red-eyed females crossed with purple-eyed males produced all red-eyed offspring

Explanations of inheritance patterns in classical genetics proceeded in two stages the first stage accounted for the transmission of genes, the second stage explained the phenotypic appearances of offspring by drawing connections between genetic make-up and phenotypic appearance [10] With regard to the first stage, geneticists identified two alleles of the purple gene the wild-type allele, designated as '+', and the mutant, purple allele, designated as '*pr*' On the basis of prior experiments, geneticists knew that the purple gene was located on chromosome II

The first stage of the explanation began with the premise, established on the basis of prior experiments (which yielded the flies used as parents in the cross discussed here), that the female parents were homozygous for the wild-type allele (+/+) and that the male parents were homozygous for the purple allele (*pr*/*pr*) Each offspring received one copy of chromosome II from its mother and one copy from its father via the processes of meiosis and gamete combination The maternally derived chromosome II in every case must have contained the wild-type allele (since both second-chromosomes of every female parent contained the wild-type allele) and the paternally derived chromosome II must have contained the purple allele (since both second-chromosomes of every male parent contained the purple allele) Hence, all offspring were heterozygous (*pr*/+) The genetic make-up of the progeny was thus explained by tracing the transmission of genes from parents to offspring, which marks the completion of the first stage of the explanation

The second stage of the explanation drew inferences about phenotypic appearances The purple allele was known to be *recessive* to wild-type This means that in relevant contexts, heterozygote flies (*pr*/+) have the same phenotypic appearance as flies homozygous for the wild-type allele Since the offspring in this cross were all heterozygous, as explained by the first stage of this explanation, and since purple is recessive to wild-type in the context of this experiment, all offspring must have the wild-type character This completed the explanation of the transmission of red-eye color from mothers to offspring

[10] I describe this experiment more fully, establish the typicality of the explanatory reasoning examined here, and discuss the larger context of investigative reasoning in which this experiment was conducted in C Kenneth Waters, "What Was Classical Genetics?," *Studies in History and Philosophy of Science*, A xxxv (2004) 783–809

The explanatory reasoning here does not depend on identifying the material make-up, mode of action, or general function of the underlying purple gene It depends only on the ideas that copies of the gene are located in chromosomes that are passed on from generation to generation and that the difference in the gene (that is, the difference between *pr* and +), whatever this difference is, causes the differences in phenotypes [11] The idea that the gene is the difference maker needs to be qualified differences in the gene cause phenotypic differences in particular genetic and environmental contexts

Difference principle differences in a gene cause uniform phenotypic differences in particular genetic and environmental contexts

The difference principle is easily applied to experimental contexts because geneticists deliberately simplified the causal situation by standardizing environmental conditions and removing genetic differences that might affect the phenotypes under study (other than the genetic differences being investigated in the experiment) Their strategy was to construct an experimental situation so that only those gene differences being investigated were causing differences in the phenotypic appearances under study

The fact that classical geneticists routinely constructed experimental situations so that the difference principle would apply undermines the notion that they naively believed that each of the phenotypic characters they studied were preformed in single genes or even gene complexes They were well aware of the fact that differences in many different genes *could* affect any given trait, and they therefore set up their experiments so the difference in a particular phenotype would be caused by a difference in one gene (or a very few genes), not many Notice that reasoning in this experimental context naturally invokes an ontological distinction between the general causes of a phenotypic trait (for example, eye color) and the cause of an actual difference in the trait (for example, the difference between red and purple eye color) Many genes are potential difference makers with respect to eye color, but in this experiment genes were controlled via breeding regimens so that only one of these genes (the purple gene) actually differed, and hence only one of these genes actually caused eye color to differ in the experiment

[11] For further argumentation on this point and the difference principle stated below, see Waters, "Genes Made Molecular," *Philosophy of Science*, LXI (1994) 163–85 Fred Gifford and Lisa Gannett also stress the importance of difference making in the context of genetics—see Gifford, "Genetic Trait," *Biology and Philosophy*, V (1990) 327–47, and Gannett, "What's in a Cause? The Pragmatic Dimensions of Genetic Explanations," *Biology and Philosophy*, XIV (1999) 349–74

This interpretation, which invokes an ontological distinction between a cause and the actual difference making cause, is born out by Morgan's own discussion of the causal relationship between genes and phenotypes Morgan, trained as an embryologist, distinguished between the point in development where geneticists typically observe a difference and the point at which the developmental process was altered to bring about the observed difference

> the study of embryology shows that every organ of the body is the end-result, the culmination of a long series of processes A change that affects any step in the process may be expected often to affect a change in end result It is the final visible effect that we see, not the point at which the effect was brought about [12]

The next quotation shows, as my analysis implies, that Morgan did not presume a simplistic one-gene one-character type view He was well aware that particular characters such as eye color are caused by many different genes He even speculated that every organ in the body might be affected by each and every gene in the organism's genome

> Suppose, for instance, to take perhaps an extreme case, all the genes are instrumental in producing each organ of the body This may only mean that they all produce chemical substances essential for the normal course of development (*ibid*)

Now, the conceptual puzzle makes its appearance

> If now one gene is changed so that it produces some substance different from that which it produced before, the end-result may be affected, and if the change affects one organ predominatingly [*sic*] it may appear the one gene alone has produced this effect In a strictly causal sense this is true, but the effect is produced only in conjunction with all the other genes In other words, they are all still contributing, as before, to the end-result which is different in so far as one of them is different (*ibid*)

This passage implies a distinction between two causal concepts It emphasizes that the end result of development is affected by, that is, caused by, all the genes working in conjunction But it also acknowledges that in another sense, in what Morgan called the "strictly causal sense," "one gene alone has produced this effect" Morgan's unanalyzed distinction corresponds to the difference between what I am calling "a cause" and what I am calling "the actual difference making cause" (or "the actual difference maker") His science entailed, as do biological sciences in general, identifying one or a few elements as

[12] Morgan, *The Theory of the Gene* (New Haven Yale, 1926), pp 305–06

the "actual cause(s)" in situations that necessarily involve many causes The problem of explaining what it means to pick out one cause among many as the actual difference maker is indeed a genuine philosophical puzzle

II WOODWARD'S MANIPULABILITY THEORY OF CAUSATION AND CAUSAL EXPLANATION

Woodward's manipulability theory of causation does not solve this puzzle, but it provides conceptual tools for constructing concepts of difference making that can explain what it means to pick out the actual difference maker His theory is based on the idea that causal relationships involve patterns of counterfactual dependencies concerning what would happen if certain properties were manipulated To say there is a causal relationship between two properties, on this account, is to say that one can change the value of one property by manipulating the other property The idea that causal relationships are those that could be used for the purposes of manipulation has been developed by a number of scientists as well as philosophers [13] Philosophers developing this idea have generally tried to construct reductive theories of causation, that is, theories that describe causal relationships in noncausal terms But Woodward, following the lead of scientists and statisticians, upon whose work he generously draws, does not aim to provide a reductive analysis of cause (as will become evident below) Instead, he assumes a basic counterfactual notion of cause, and then uses ideas about manipulations developed in the scientific and statistical literatures to construct a general theory about causal relationships and an account of various ways scientist reason about them

Although Woodward's account includes interesting ideas about causal inference and explanation, I will not describe them here because my argumentation centers on ontological, not epistemological, considerations In addition, I will not review how Woodward responds to various criticisms of manipulability theories But it is worth pointing out that since Woodward's theory is not reductive, it avoids difficulties that vex traditional manipulability theories of causation [14]

[13] Prominent philosophical accounts include R G Collingwood, *An Essay on Metaphysics* (New York Oxford, 1940), D Gasking, "Causation and Recipes," *Mind*, LXIV (1955) 479–87, Georg Henrik von Wright, *Explanation and Understanding* (Ithaca, NY Cornell, 1971), and Peter Menzies and Huw Price, "Causation as a Secondary Quality," *British Journal for the Philosophy of Science*, XLIV (1993) 187–203 Accounts developed by nonphilosophers include Thomas D Cook and Donald T Campbell, *Quasi-Experimentation Design and Analysis Issues for Field Settings* (Boston Houghton Mifflin, 1979), and Judea Pearl, *Causality* (New York Cambridge, 2000)

[14] For example, a common criticism of manipulability accounts of causation is that by reducing statements about causal relationships to statements about the possibilities

Woodward motivates the stress on manipulability by pointing out that human interest in causal relationships began with practical interests in manipulation and control This motivation certainly resonates with the history of biology A central theme in the history of experimental biology (and genetics in particular) is that persistent interest in manipulating and controlling life, both on the part of scientists and those providing scientists with resources, has exerted a tremendous impact on the course of research Even in the cases of evolutionary biology and ecology, which are often pursued in nonexperimental venues, research has often been inspired, even driven, by human interests in manipulation and control Hence, even though Woodward developed his theory of causation largely in the context of physics and economics, the motivation for causal reasoning made salient by his theory is very relevant to causal reasoning in biology

Woodward, following the practice of many scientists and statisticians, takes the relata of causal relationships to be properties (rather than events) To say that there is a causal relationship between *X* and *Y*, on his account, is to say that *X* and *Y* are properties and that manipulating one property would change the other Woodward conceives of causal relationships as relationships between variables, so he would express this idea by saying that the value of the effect variable, say *Y*, can be changed by manipulating the value of a causal variable, say *X*. Variables in this context designate properties that take on two or more mutually exclusive values The variable of temperature, for example, designates a magnitude that can take values such as 20 degrees or 100 degrees and the phenotypic variable of eye color designates a property that can take on values such as red or purple These values are possessed by or instantiated in particular entities So, for example, the water in a kettle might be 20 degrees or the eyes of a fruit fly might be red So, according to

---

of changing effects by manipulating causes, they seem to assume that causes can in fact be manipulated But Woodward's theory makes no such assumption, rather, it assumes that there is a truth to the matter of what would happen if it were technologically (or otherwise) possible to manipulate the relevant properties and the properties were manipulated On Woodward's theory, nothing prevents reading both instances of 'were' in this sentence as possibly counter-to-fact A critic might respond that with respect to some phenomena, such as quantum phenomena, the very notion of manipulation ceases to make sense If this is indeed the case, then one might defend a manipulability theory like Woodward's on the ground that quantum phenomena are (therefore) not causal See note 16 for further discussion of how Woodward's theory avoids, or critics might say "postpones," difficulties confronting reductive theories of causation

Woodward's theory, causal relationships are relationships between values of distinct properties that can be represented as values of different variables

Roughly speaking, Woodward writes, a causal relationship exists between two variables, X and Y, when

> for at least some individuals there is a possible manipulation of some value of X that they possess which, given other appropriate conditions (perhaps including manipulations that fix other variables distinct from X at certain values), will change the value of Y or the probability distribution of Y for those individuals [15]

The inclusion of the causal term 'change' in the above passage reflects the fact, already mentioned, that Woodward's theory is not reductive It does not account for cause in noncausal terms [16] It is a nonreductive, counterfactual theory, potentially consistent with different reductive theories, including counterfactual theories such as David Lewis's While some metaphysicians will not be satisfied until they have a reductive theory, perhaps they would be willing to concede that Woodward's theory provides a useful way to think about causes in the meantime In any case, this is the modest premise upon which the analysis that follows is based

Two technical concepts are important for understanding the gist of Woodward's theory intervention and invariance When Woodward refers to "a possible manipulation of X" he has in mind a special kind of manipulation It is not the case that any manipulation of the value of X that subsequently changes the value of Y implies that that X causes Y For example, a manipulation of the thermometer reading brought about by decreasing temperature would change the phase state of water, but this does not mean that thermometer readings cause water to freeze To say that the thermometer reading causes water to freeze is to say that a special kind of manipulation of the thermometer reading, a manipulation that did not independently affect other causal variables such as temperature, would change the

[15] Woodward, *Making Things Happen* (New York Oxford, 2003), p 40

[16] Woodward's theory, because it is not reductive, cannot solve certain fundamental puzzles about causation, including the problem of coincidence Because the theory is not reductive, it contains no ultimate answer to those who would ask of the quoted passage but what distinguishes the claim that the intervention in X changes the value of Y from the claim that the intervention was merely a coincidence? Woodward's theory provides answers that are useful for understanding how to clarify causal ideas or empirically test them, but its answers presuppose a basic counterfactual notion of cause that is not itself explained Insofar as puzzles relate directly to this notion, Woodward's theory will not solve them But this does not imply that his theory is false, rather, it reflects the fact that his theory is not reductive

phase state from liquid to solid The technical term for the relevant kind of manipulation is 'intervention' A manipulation counts as an *intervention on X with respect to Y* if and only if the manipulation changes the value of X without changing, independently of the change in value of X, the value of any causes of Y[17] 'X causes Y' means that an intervention on X with respect to Y would change the value of Y Notice that Woodward's account does not require that *all* interventions on X would change Y, and this brings us to the second technical concept of his theory invariance

On Woodward's account, causal relationships are not necessarily universal X can cause Y even if intervening on the value of X within many ranges of the variables X and Y would not change the value of Y For example, there is a causal relationship between temperature and the phase state of water even though increasing the temperature of water from 22° C to 95° C will not change its phase state In addition, X can cause Y even if intervening on the value of X has no effect on the value of Y when the value of certain variables not identified in the expression 'X causes Y' is outside some range For example, the claim that raising the temperature of water from 22° C to 100° C causes water to boil is not contradicted by the fact that increasing the temperature of water from 22° C to 100° C will not change the phase state of water when the atmospheric pressure is significantly greater than standard atmospheric pressure An advantage of Woodward's theory is that it accounts for the intuitively plausible idea that scientists typically explain phenomena by identifying causal dependencies that do not obtain under all conditions The thermodynamical explanation of why a flask of water boiled over the flame of a Bunsen burner does not depend on identifying a totally invariant relationship Identifying a change in water temperature as the cause of the change in phase state under the laboratory conditions suffices

Woodward says that scientists identify genuine causal relationships even though these relationships (for example, the ones specified in textbooks) are almost always "sensitive" to the values of variables that figure into the specified relationships and/or to the values of variables that do not figure into the specified relationships According to his account, the expression 'X causes Y' just means there are *at least two* different values of X, x_1 and x_2, and *at least one* set of values for the variables specified in the expression and the variables not specified in the expression, such that if one intervened to change the value of variable X from x_1 to x_2, the value of Y would change The

[17] This is my description of Woodward's concept of intervention He carefully sets out a technical definition that deals with subtleties that can be ignored for present purposes

invariance of a causal relationship concerns ranges of values under which the relationship holds Since causal relationships can be sensitive to the values of many different variables, and since the sensitivity with respect to one variable can depend on the values of other variables, the relevant 'ranges of values' refers to a space (or set of spaces) in a n-dimensional array of variables (where n is the number of variables to which the relationship X causes Y is sensitive) Hence, it is more accurate to speak of "spaces of invariance" than "a range of invariance "

According to this account, causal explanations in science cite causal relationships exhibiting various amounts of invariance Any such relationship must be at least *minimally invariant* in the sense that it must hold for at least two different values of the causal variable (a very weak condition) And it may be *maximally invariant* in the sense that it holds for all values of all variables whether they explicitly figure into the specification of the causal relationship or not (a very strong condition) But scientific explanations typically cite causal relationships that fall somewhere between these two extremes Woodward reserves the term 'law' for maximally invariant causal relationships A virtue of his account is that it explains what is special about scientific laws (that is, maximal invariance) without implying that this special feature is required for scientific explanation

Woodward's theory provides a natural framework for expressing the causal relationships identified by classical geneticists The difference principle, discussed in section I, can be reformulated in terms of the manipulability theory

> Differences in a gene cause uniform phenotypic differences in particular genetic and environmental contexts
>
> ⇓
>
> For at least some organisms, a manipulation of the form of a gene they possessed as a zygote, given other appropriate conditions (such as genetic background and environmental conditions), would change the phenotypes of those individuals

This reformulation makes salient features of the causal relationship expressed by the difference principle that were crucial for the practice of classical genetics Experimentation in classical genetics entailed controlling the genetic background and environmental conditions Geneticists bred model organisms to "clean up" their genetic backgrounds so phenotypic effects of particular mutations would not be masked They also standardized environmental conditions in the laboratory These efforts resulted in regular and repeatable patterns of inheritance Woodward's theory emphasizes the fact that causal

relationships, such as those expressed by the difference principle, are context sensitive

III A LACUNA IN THE MANIPULABILITY THEORY

Woodward's theory provides insights concerning many philosophical questions about causation and causal explanations, but it does not solve the problem about what it means to pick out the cause that makes an actual difference This lacuna can be illustrated by returning to the experiment discussed in section I In this experiment, the purple gene (the gene that takes on one of two contrasting values, + or *pr*) was identified as the cause of red eye color in the offspring even though this phenotypic end-result was known to depend on a number of different genes and environmental factors According to Woodward's manipulability theory, the purple gene was indeed a cause because the eye color of offspring could have been manipulated by intervening on this gene (for example, by changing the + allele in heterozygous offspring embryos to a *pr* allele) This idea is an *instantiation of the difference principle*

> For the offspring in the experiment, a manipulation of the maternally derived gene from + to *pr*, given other appropriate conditions (such as the genetic background and environment), would change the phenotypes of the offspring from red-eyed to purple-eyed

But Morgan and his collaborators knew that eye color could be manipulated by intervening on a number of different genes (by 1919 they had identified mutations in at least 16 different genes located at separate loci that affected eye color) For example, just as eye color of the offspring in the experiment could have been manipulated by intervening on the purple gene, eye color could have been manipulated by intervening on the vermilion gene (for example, by changing the wild-type alleles in the offspring, which were homozygous for the vermilion wild-type allele, to the vermilion mutant allele) Hence, according to Woodward's theory, the vermilion gene was just as much a cause of red eye color in these offspring as was the purple gene [18] So Woodward's theory clarifies what it means to identify the purple gene as a cause of red eyes, but it implies that the vermilion gene is just as much a cause

[18] Readers might think this problem could be easily resolved since the vermilion mutation results in a different phenotype than that the purple mutation But there are other examples where different mutations (or different combinations of mutations) bring about phenotypic effects that are indistinguishable from one another My account will explain why, in even these cases, it can make sense to say that one gene and not the other (or one set of genes and not the other set) is the actual difference maker

Woodward tries to address the issue of why some causal factors are more significant in particular cases than others when he discusses token causation, where the problem is to pick out "*the* cause" of a single event [19] But Woodward's analysis of token causation, which he presents only tentatively, does not provide a basis for elucidating the "strict" sense of cause that Morgan alluded to when he wrote "it may appear the one gene alone has produced this effect In a strictly causal sense this is true, but the effect is produced only in conjunction with all the other genes "[20]

IV THE CONCEPT OF THE ACTUAL DIFFERENCE MAKER

The concept missing from Woodward's theory of cause and causal explanation, and indeed missing from the philosophical literature in general, is that of an actual difference making cause This concept captures Morgan's "strict" sense of cause, the sense in which the purple gene, and not the vermilion gene, was "the cause" of eye color in the experiment Causation is often thought of as difference making What is typically overlooked, however, is the fact that in order for there to be a difference maker there must be a difference And in order for there to be an *actual* difference maker, there must be an *actual* difference This raises the question an actual difference among what things? The actual difference might be a difference among different entities, that is, among different actual entities, or it might be a difference in the same entity at different actual times

The first step in identifying an actual difference making cause, or an actual difference maker, is to specify an actual difference among actual entities (or an actual difference in the same entity at different actual times) If there is no actual difference, then there is no actual difference maker In the classical genetics experiment discussed earlier, there was an actual difference, it was a difference in eye color among mother, father, and offspring flies in the experiment The actual difference maker was the purple gene because it was the cause that actually differed and whose difference led to the actual difference in eye color (red versus purple) The other genes, including the vermilion gene, were causes, but they were not actual difference

[19] See Woodward, *Making Things Happen*, pp 74–86

[20] Quoted from Morgan, *The Theory of the Gene*, pp 305–06 Woodward takes readers through a complicated attempt to deal with examples of actual causation, which fails because he tries to identify 'the' in the context of singleton causal events I discuss the problem with such attempts later in this paper My aim here is to show that the problem of picking out *the actual difference making cause* arises within his theoretical framework I prefer to offer a solution within his framework that works rather than present a detailed argument about why his tentative analysis of a similar problem will not do the job

makers in the experimental population of flies The other genes that cause differences in eye color could not have been actual difference makers in the experimental population for the simple reason that they did not actually differ and hence it was not differences in them that brought about the actual differences in the population

The concept of the actual difference maker can be specified as follows

X is *the actual difference maker with respect to Y in population p* if and only if

(i) *X* causes *Y* (in the sense of Woodward's manipulability theory)
(ii) The value of *Y* actually varies among individuals in *p*
(iii) The relationship expressed by '*X* causes *Y*' is invariant with respect to the variables that actually vary in *p* (over the spaces of values those variables actually take in *p*)
(iv) Actual variation in the value of *X* fully accounts for the actual variation of *Y* values in population *p* (via the relationship *X* causes *Y*)

Notice that this concept applies only to causes in actual populations It makes no sense to identify something as the actual difference maker without identifying either a population with at least two entities that actually differ with respect to the effect variable *Y* or a population of one or more entities that exhibit different *Y* values at different actual times It is worth stressing that the actual effect is not a single property in a single token, it is a difference of a property in a population [21]

Woodward's theory provides the tools to specify condition (iv) more fully Consider what it means to say, for instance, that variations in the purple gene fully accounted for the actual variation in red versus purple eye color in the experimental population It means that the phenotypic value was a function of the genotypic value in this population and that the generalization '*X* causes *Y*' produced the correlation Or more concretely (a) there was an actual difference in eye color among flies in the population, (b) flies in the population with the same genotypic values (with respect to this gene) had the same eye color, and (c) if the flies in the experiment had been intervened on such that they all had the same genotypic value (either +/+, +/*pr* or *pr*/*pr*), then this difference in eye color would not have existed Or more generally

Actual variation in the value of X fully accounts for the actual variation of Y values in population p (via the generalization '*X* causes *Y*') if and only if conditions (i) – (iii) above hold and

[21] I am using 'population' in a statistical rather than biological sense a population is either a set of entities or possibly the set of instances of an entity at different times

(a) Individuals with the same X values in p have the same Y values
(b) An intervention on X with respect to Y that changed the X value of all individuals in p to the X value that one and the same individual had sans intervention would change Y values in p such that they no longer differed
(c) There is no variable Z, distinct from X, such that an intervention on Z with respect to Y that changed Z values in one or more individuals in p to the Z value that one of the individuals had sans intervention would change Y values in p

V SOLVING THE PUZZLE ABOUT WHAT IT MEANS FOR ONE CAUSE AMONG MANY TO BE THE ACTUAL DIFFERENCE MAKER

The concept of the actual difference maker clarifies the idea that there is an ontological difference between being a cause and being the cause that actually made a difference Being a cause only entails that something is a potential difference maker, not the actual difference maker To be a *potential difference maker*, that is, to be a causal variable, it suffices to satisfy the counterfactual patterns specified by the manipulability theory It does not matter whether the causal variable actually varies in any actual population and whether this variation brings about actual differences But to be the actual difference making cause of an actual difference in a population, the value of the variable must actually differ and this variation must bring about the actual differences among the entities in the population This is how to make sense of the idea, invoked by Morgan, that it is one thing to be a cause and another thing to be the cause strictly speaking Seeking the one cause among many that actually made the difference is no fool's errand It entails picking out the actual difference maker among potential difference makers in an actual population The philosophical puzzle with which we began is solved

In fact, the puzzle with which we began does not seem to be much of a puzzle at all Picking out the actual difference making causes among many causes simply involves picking out the causes that actually differ and whose differences account for the actual differences of interest It turns out that what makes the conceptual situation puzzling is not so much a puzzling concept, but puzzling applications, such as the application with which I began, the lighting of the match To apply the concept to this situation, to say that Mary's striking the match is the actual difference making cause whereas the presence of oxygen was merely a cause, requires identifying an actual difference in an actual population If Mary took the match from a box, then the matches contained in the box were an actual population and the difference between having lit versus not having lit was an actual difference With respect to this actual effect (that is,

the actual difference in this actual population), it is obvious that Mary's striking the match was the actual difference making cause and that the presence of oxygen was merely a cause The presence of oxygen was not the cause that made the actual difference in this actual population of matches because, after all, oxygen was present for all the matches [22]

Before proceeding to the next section, which considers more complex causal situations, it is helpful to consider two possible objections, one concerning whether my resolution of the puzzle is ontological, and the other concerning my use of the terms 'actual' and 'potential' Readers might question whether this account identifies an ontological feature After all, the selection of a difference making cause depends on the prior identification of a population Can populations be identified, or specified, independently of our interest? If not, then applying the distinction between potential and actual difference makers seems to depend on more than ontology But this line of questioning conflates the identification of a cause with the specification of an effect Of course, being interested in one effect rather than another depends on more than the ontology of causal processes Our interests lead us to inquire about the causes of some things and not others The question here is whether the cause(s) of an effect are fixed by ontology This account of difference making takes an effect to be an actual difference in an actual population My claim is that the subset of causes that are the actual difference makers of such an effect is fixed entirely by the ontology of causation in the actual population, not by a mix of ontology and interests as recent writers have maintained [23] This is not to say that the selection of some causes among many is not based on interest or that it is based on a universal interest to pick out actual difference makers But interests often are based on an interest to pick out actual difference makers in one or another actual population, and when they are, the issue reduces to ontology the variety of epistemic factors detailed by Hart and Honoré are irrelevant and the features of real populations specified in this analysis determine which causes are the actual difference makers The error of this possible objection is to infer from the fact that the selection of effect in an epistemic context involves

[22] This analysis might be extended to singleton cases as well If there were no other matches, then one could consider the effect to be the difference in the population made up of instances of the match before, during, and immediately after it was struck

[23] Even Jonathan Schaffer, who comes closest to offering an ontological account, appeals to pragmatic factors to resolve the selection problem See Schaffer, "Contrastive Causation," *Philosophical Review*, CXIV (2005) 297–328

pragmatics to the mistaken idea that what counts as the cause of that effect must also depend on pragmatics My point is that once the effect is fully specified as an actual difference in a real population, the issue of which causes are the actual difference makers is an ontological one

Readers might also object that my use of the terms 'actual' and 'potential' is misleading because all causes, even what I call potential difference making causes, are actual in the sense that the actual differences causally depend on them It is easiest to clarify this point by using the terminology of variables An actual effect, on my account, is the actual difference in values of an effect variable in an actual population In the situations examined thus far, one causal variable plays a differentiating role in the sense that actual differences in its value explain the actual differences in values of the effect variable Other causal variables do not play this role because they exhibit uniform values in the actual population They are, however, potential difference makers because if, counter to fact, their values varied in the population then this variation could result in an increase in the differentiation of the values of the effect variable In addition, however, a counter-to-fact variation in the values of a potential difference making variable might result in a reduction of the differentiation of the values of the effect variable (compared to the values of this variable in the actual population) Hence, the difference in values of the effect variable in the actual population depends not just on the different values taken by what I am calling the actual difference making variable, but also on the uniform values taken by what I am calling the potential difference making variables So, one might say all causal variables are "actual" difference makers in the sense that the actual differences depend on all their values Nevertheless, the fact that the values of some causal variables actually differ and the values of other variables do not is a feature of the causal structure of an actual population I am choosing to reserve 'actual' to designate the causal variables whose values actually differ and 'potential' to designate the causal variables whose values do not actually differ This terminology comes very naturally in the contexts of causal reasoning in the special sciences, which will become evident when we turn our attention to the complex biological processes such as development

In summary, the key to solving the conceptual puzzle, the key to understanding what it means to select an actual difference making cause from a plethora of mere causes, is to see that the selection involves picking out the causal property, differences in which produced an actual difference in a population

VI THE CONCEPT OF AN ACTUAL DIFFERENCE MAKER

In many biological situations, there is not just one actual difference maker, there are many That is, actual variations in two or more variables in an actual population account for actual differences in an effect variable In such situations, the operative concept is *an* actual difference maker, not *the* actual difference maker

X is *an actual difference maker with respect to Y in population p* if and only if

(i) *X* causes *Y*
(ii) The value of *Y* actually varies among individuals in *p*
(iii) The relationship *X* causes *Y* is invariant over at least parts of the space(s) of values that other variables actually take in *p* (In other words, it is invariant with respect to a portion of the combinations of values the variables actually take in *p*)
(iv) Actual variation in the value of *X* partially accounts for the actual variation of *Y* values in population *p* (via the relationship *X* causes *Y*)

As in the case of the actual difference maker concept, the fourth condition in this construction can be specified by using Woodward's concept of intervention

X partially accounts for the actual variation of Y values in population p (via the relationship *X* causes *Y*) if and only if conditions (i) – (iii) above are satisfied and

An intervention on *X* with respect to *Y* that changed the *X* values in one or more individuals in *p* to the *X* value that one of the individuals had sans intervention would change *Y* values in *p*

The statistical relationship between *X* and *Y* values in the case of *an* actual difference maker is far messier than in the case of *the* actual difference maker In the latter case, the value of *Y* is a function of *X*. But in the case of *an* actual difference maker, the individuals in the population with the same *X* value might have different *Y* values

The causal relationship underlying an actual difference maker claim is also more complicated In the case of the actual difference maker, eliminating the variation in *X* values (by intervening on *X* with respect to *Y* in *p*) would eliminate the variation in *Y* values In the case of an actual difference maker, eliminating the variation in *X* values (by intervening on *X* with respect to *Y* in *p*) does not necessarily eliminate the variation in *Y* values It only changes the variation in *Y* values In fact, eliminating the variation in the value of *an* actual difference maker might actually increase, rather

than decrease, variation in the value of Y in p So, it would be more accurate to use the term "an actual difference maker" with "an actual difference changer" and to specify that an actual different changer might be "an actual difference increaser" or "an actual difference reducer" in the population This fine conceptual point has important implications in real science, for example, in quantitative genetics

Quantitative genetics involves a number of subtleties of actual difference making While analyzing these subtleties is a topic for another paper, some intuitively plausible connections between actual difference making relationships in different populations can be specified quickly Suppose there are three populations, p_1, p_2, and p_3, that p_1 and p_2 are mutually exclusive, and that p_1 is a proper subset of p_3 If X is the actual difference maker with respect to Y in p_1, it is not necessarily the actual difference maker with respect to Y in p_2 In fact, it is not necessarily an actual difference maker with respect to Y in p_2 But if X is the actual difference maker in population p_1, then X is at least an actual difference maker in p_3 Furthermore, if X is an actual difference maker in p_1, then it is not necessarily an actual difference maker in p_2, but it is necessarily an actual difference maker in p_3

VII DOES THE EXISTENCE OF MULTIPLE CAUSES INDICATE THAT THERE IS NO ONTOLOGICAL BASIS FOR CENTERING ATTENTION ON GENES AND DNA?

My solution to the conceptual puzzle about the difference between what it means to be a cause and what it means to make an actual difference reveals the fallacy of causal parity arguments The fact that there are many causes does imply that there is ontological parity among them DNA is one cause among many, but in some contexts it might be the actual difference maker while other causes are only potential difference makers My aim in this section is to identify situations in which DNA is an ontologically distinctive cause, and to clarify the nature of its causal distinctiveness in these situations It turns out that understanding why DNA is a distinctive cause requires invoking a concept of causal specificity as well as concepts of actual difference making

As readers will see, the situation I identify is a small part of the developmental process and hence my analysis does not provide the basis for making sweeping claims against all theses of causal parity But by challenging the basic logic of causal parity arguments, and by showing that parity breaks down in certain contexts, my account reveals important clues for understanding why

so much research attention in developmental biology is centered on DNA [24]

Recall that the first step for identifying the (or an) actual difference maker is identifying an actual difference in an actual population An important actual difference in cells is the difference in the nucleotide sequences of RNA molecules synthesized in a cell This is not the only important actual difference in a cell, but it is an important difference Consider a relatively simple example, the population of RNA molecules synthesized in a bacterium Synthesis of any RNA molecule in a bacterium depends on the presence of an activated gene (a segment of nucleotides in a DNA molecule), RNA polymerase, and several additional accessory proteins All of these molecules are causes, so they are all potential difference makers Intervening on any of the activated DNA segments could alter the sequence of nucleotides in the resulting RNA molecules, and intervening on any of the others (for example, RNA polymerase) could stop RNA synthesis altogether They are all potential difference makers

But only the activated DNA segments (the genes) are actual difference makers of RNA sequences in a bacterium Let G be the variable representing the nucleotide sequences in the activated genes, P be the variable representing the structure of RNA polymerase molecules, $P_1 - P_n$ designate variables representing structures for each of the accessory proteins, and R designate the variable representing the nucleotide sequences in the RNA molecules Each of the variables G, P, P_1, , P_n represent causes of R, so each satisfies the first condition for being the actual difference maker But among these potential difference makers, only G takes on different values Hence, only G satisfies the second condition for being the actual difference maker Since the other variables do not take on different values, the causal relationship between G and R is invariant with respect to the differences in the bacterial cell So the third condition for being the actual difference maker is satisfied Furthermore, actual variation in the value of G completely accounts for actual variation in R values in the sense spelled out in the fourth condition of the actual difference maker concept That is, individual RNA molecules synthesized from the same

[24] Other aspects relevant to a fuller account of this issue are in Waters, "What Was Classical Genetics?," and Waters, "A Pluralist Interpretation of Gene-Centered Biology," in Stephen Kellert, Helen Longino, and Waters, eds , *Scientific Pluralism Minnesota Studies in the Philosophy of Science, Volume XIX* (Minneapolis Minnesota UP, 2006), pp 190–214, and Waters, "Beyond Theoretical Reductionism and Layer-Cake Antireductionism How DNA Retooled Genetics and Transformed Biological Practice," in Michael Ruse and David Hull, eds , *Oxford Handbook to the Philosophy of Biology* (forthcoming)

gene have the same R values and if the activated genes in the cell had been intervened on so that each had the same G value (the same nucleotide sequence), then the synthesized RNA molecules would have had the same R value as well (the same nucleotide sequence) With respect to the differences in the linear sequences of the population of RNA molecules in a bacterial cell, current evidence indicates that DNA is the actual difference making cause

RNA synthesis in eukaryotic cells is more complicated It involves three different kinds of RNA polymerase molecules and many dozens of accessory proteins The different kinds of polymerase molecules are involved with the synthesis of different classes of RNA molecules (for example, RNA polymerase I catalyzes the synthesis of rRNA and RNA polymerase II catalyzes the synthesis of mRNA) Presumably, different accessory molecules are also associated with the synthesis of different RNA molecules In this situation, DNA is not the only actual difference maker because it is not the only cause that actually differs among the synthesis events in a cell For example, consider the difference in linear structures of the population of mRNA and rRNA molecules in a eukaryotic cell The genes involved in the syntheses differ, but so do the RNA polymerases Intervening on polymerase II could stop synthesis of mRNA altogether, and this would affect the structures of the resulting RNA molecules Intervening on polymerase I could have a parallel effect, as could intervening on accessory molecules Hence, DNA is *an* actual difference maker, not *the* actual difference maker RNA polymerase and accessory proteins could also be actual difference makers Nevertheless, many other background constituents necessary for the synthesis to occur (for example, water) do not differ So the logic of the parity argument does not hold But this does not fully resolve the issue about parity because the different actual difference making causes, say genes versus polymerases, do not seem to be on an ontological par

Further resolution of the alleged parity requires an additional causal concept, one of causal specificity The concept I have in mind is similar to Lewis's concept of causal influence, and can be sketched by example [25] DNA is a *specific difference maker* in the sense that different changes in the sequence of nucleotides in DNA would change the linear sequence in RNA molecules in many different and very specific ways RNA polymerase does not have this specificity Interven-

[25] Lewis, "Causation as Influence," this JOURNAL, XCVII, 4 (April 2000) 182–97 Conversations with Chris Hitchcock and Jim Woodward helped me develop this idea Woodward pointed out that my ideas about causal specificity had a lot in common with Lewis's ideas about causal influence

ing on RNA polymerase might slow down or stop synthesis of a broad class of RNA molecules, but it is not the case that many different kinds of interventions on RNA polymerase would change the linear sequence in RNA molecules in many different and very specific ways This shows that DNA is a causally specific potential difference maker The fact that many such differences in DNA do actually exist and these differences actually explain the specific differences among RNA molecules indicates that DNA is the causally specific actual difference maker with respect to the population of RNA molecules first synthesized in eukaryotic cells

As readers familiar with contemporary molecular biology know, RNA production in eukaryotes is often complicated by processes that change RNA molecules after they are synthesized If we consider the wider population of RNA molecules in a eukaryotic cell (not the subset of those first synthesized), then we need to consider processes such as RNA splicing RNA splicing changes the linear sequence of an RNA molecule by removing certain internal segments in the RNA molecule and "splicing" the remaining pieces together RNA splicing is causally specific Often, one and the same RNA molecule is spliced in different ways by different splicing agents In these situations, the splicing agents are causally specific actual difference makers They actually differ and the actual differences lead to specific changes in the nucleotide sequence of the mature DNA molecule In the production of these molecules, DNA is not the only causally specific actual difference maker [26] We might say that the structure of these mature RNA molecules are determined by DNA (that is, the activated gene) and the splicing agent In the case of RNA synthesis in eukaryotes the problem of causal parity is resolved, not by showing that DNA is the only causally specific actual difference maker, but by showing that it is one of several [27]

Explanations in molecular biology do not rest on the premise that genes are the only causally specific actual difference makers with respect to the linear sequence of nucleotides in primary RNA, processed RNA, and polypeptides, but they do deny the idea that all causal factors involved in the production of RNA and polypeptide

[26] But as Marcel Weber has pointed out to me, DNA has a greater degree of causal specificity than the other actual difference makers

[27] I have limited discussion here to one of the apparently several kinds of causally specific actual difference makers involved with the synthesis of RNA in eukaryotes Discussing more examples would simply belabor my philosophical point that not all causes are on a par because some are causally specific actual difference makers (for example, genes and splicing agents) and others are not (for example, RNA polymerases I and III)

molecules are on a causal par Such a parity thesis is false If true, it would indicate that emphasizing DNA and splicing agents in explanations of mRNA synthesis has no ontological basis But current evidence indicates otherwise The initial synthesis of RNA in prokaryotes and eukaryotes involves many causes including polymerases and accessory proteins, but only DNA is the causally specific actual difference maker Apparently that is the end of the story for prokaryotes In eukaryotes, some very important RNA molecules (and perhaps many) are differentially processed, so splicing agents are also causally specific actual difference makers The linear sequence of nucleotides in mRNA is determined by the linear sequence of nucleotides in DNA, and in cases of differential splicing also by other molecules such as alternative splicing agents [28]

VIII WHAT KIND OF CAUSAL GENERALITY MATTERS IN BIOLOGY?

Biologists seek a special kind of generality in causal statements Whereas physical scientists often try to establish generalizations that express causal relationships that hold over conditions that have not and will not be actualized (to their knowledge), biologists typically care only about whether a relationship holds under conditions actualized in organisms and their environments (or in the laboratory) [29] Biologists do not care whether causal generalizations linking sequences in mRNA to sequences in polypeptides in the causal processes of polypeptide synthesis would still hold if the cellular conditions were quite different than cellular conditions actually are For instance, if the form of tRNA molecules were different, then an mRNA molecule that now causes a particular amino acid sequence would cause a very different amino acid sequence The fact that different tRNA could render the current causal generalizations about polypeptide synthesis false does not matter to biologists unless different tRNA has actually existed or is likely to exist actually in the future We are now in a position to clarify the kind of causal generality that matters in biology

The difference can be articulated in terms of invariance space Recall that a causal relationship is not necessarily maximally invariant, it can hold under some combinations of variable values and not under others I introduced the term *invariance space* to refer to the set of value combinations under which a causal relationship holds The observation mentioned above can be expressed this way physical

[28] See previous footnote

[29] This claim needs to be qualified Biologists interests in manipulating causal processes in organisms sometimes motivates them to seek causal generalizations that hold under conditions that could be artificially brought about in organisms and their environments See final paragraph in this section

scientists often aim to identify causal relationships whose invariance space includes combinations of variable values that are not, as far as humans know, ever actualized The fact that a portion of invariance space is never occupied does not seem to render the question of whether the relationship holds in that portion of invariance space less interesting or important in fundamental physics But in the case of biology, the fact that a combination of variable values is never actualized does render the question of whether a relationship holds under that set of variable values less interesting and perhaps even altogether irrelevant Consider the variable of tRNA forms If the values of this variable were different than they actually are, that is, if actual tRNA molecules had different forms than they actually have, then the causal generalizations associating sequences in RNA and sequences in polypeptides could be strikingly different And biologists are intensely interested in whether the values of this variable actually differ in certain mitochondria and chloroplast organelles But if the values do not actually differ, then it does not much matter to biologists whether the invariance space of their generalizations include the nonactualized values

The reason biologists do not care whether the invariance space of a causal relationship includes combinations of variable values that are not actualized is because they are much more interested in the actual than the possible Biologists are more interested in explaining actual differences than possible differences, and explaining actual changes over developmental and evolutionary time than possible differences This is why they are more interested in identifying actual difference makers than potential difference makers [30] This also explains why they are satisfied with highly sensitive causal generalizations about actual difference makers Whether a property is an actual difference maker depends on whether it makes the difference under the conditions that actually obtain in the actual biological world Whether it would make that difference if the conditions were different is irrelevant to understanding the difference the property makes in the actual world Hence, seeking causal generalizations that are invariant over combinations of variable values that have not existed will not, directly at least, further the project of identifying the causal relationships relevant to explaining the actual differences in nature or the actual changes in complex biological processes such as ontological development

[30] I thank Michael Strevens for emphasizing the idea that biologists are interested in actual difference making because they are interested in the actual, and not the other way around

The last claim is qualified ("directly at least") because biologists' interests include manipulating biological entities This interest stems not just from the motivation to improve conditions for humans (for example, to cure diseases, increase agricultural yields, enhance the aesthetic value of plants and animals), but also from the motivation to investigate processes and mechanisms in living systems A chief strategy of investigation in biology is to manipulate basic biological processes and mechanisms to learn how they work [31] Hence, it is relevant whether causal generalizations will hold under conditions that could be made actual even if they are not currently actual With this important qualification in mind, we can say that with respect to the goal of explaining the processes of life such as physiological functioning, ontological development, and evolutionary change, it is irrelevant whether the specified causal relationships extend to invariance spaces that include nonactualized combinations of variable values This explains why biologists' interest in the actual leads them to be perfectly satisfied with causal generalizations that are highly sensitive even if physical scientists are not

IX CONCLUSION

The philosophical literature on causation includes attempts to bring one or another theory of causation into accord with intuitions about what was "the cause" of an event The literature seems to take it for granted that our intuitions are distinguishing causes from noncauses But the analysis of this paper suggests that intuitions about what was "the cause" in an actual situation might stem from a natural inclination to pick out actual difference making causes, not an inclination to pick out causes simpliciter Psychological research on causal reasoning in humans supports this idea Recent empirical findings indicate that which factors are picked out as causes is indeed influenced by whether a candidate causal variable actually takes different values in the entities under observation [32] The possible relevance of actual variation in the values of causal variables goes unnoticed in the philosophical literature when effects such as the lighting of a match are described as a singular actual outcome in a singleton rather than as a difference in actual outcomes in a multiplicity Hence, the ontological contexts in which the actual difference making causes are actually making a difference are removed from consideration

[31] See Waters, "What Was Classical Genetics?," Waters, "A Pluralist Interpretation of Gene-Centered Biology," and Waters, "Beyond Theoretical Reductionism and Layer-Cake Antireductionism How DNA Retooled Genetics and Transformed Biological Practice "

[32] Patricia W Cheng, "From Covariation to Causation A Causal Power Theory," *Psychological Review*, CIV, 2 (1997) 367–405

and it is therefore impossible to identify which cause is the actual difference making cause Our intuitive judgments about what cause is making an actual difference ends up being based on tacit hunches about what is missing from the descriptions If biologists were similarly limited to considering the causal synthesis of a single polypeptide molecule, they would have no basis for saying that the polypeptide's linear sequence was determined by DNA, and not by RNA polymerase In fact, if restricted to considering a single instance (or a population of identical outcomes), it might appear that DNA was merely scaffolding for the synthesis of RNA The causal distinctiveness of DNA is in the population It is only in the context where polypeptide molecules with different amino acid sequences are being synthesized that it makes sense for biologists to say that DNA is not on a causal par with many of the other molecules that play causally necessary roles in the synthesis of RNA and polypeptides Likewise, I suspect, focusing attention on singleton situations about a single lighting of a match, a single breaking of a vase, or the single catastrophic dropping of a boulder obscures important features of causation Much light could be shed on causal reasoning by shifting attention to causes in populations

C KENNETH WATERS

University of Minnesota

CENTRAL LIBRARY
UNIVERSITY
OF
CALCUTTA
TAGORE CAMPUS

REVIEW ESSAYS

PREDICATION AND TRUTH

Truth and Predication DONALD DAVIDSON Cambridge Harvard University Press, 2005 x, 180 p Cloth $29 95

Donald Davidson had nearly completed this book when he died unexpectedly at age 86 The book had been accepted for publication He had received comments from Charles Parsons, Jim Higginbotham, and the present reviewer After Davidson's death, his widow Marcia Cavell, helped by colleagues, made conservative changes based on notes or marginalia that Davidson had written Where the intent of such notes was not completely clear, she entered them in bracketed footnotes Although there is some reference to Parsons's remarks in such footnotes, it appears that Davidson had not made significant revisions in response to any of the reviewers' comments Cavell judges that he would have made more revisions So it is impossible to know what a completed book would have looked like Still, the book reads like a work conceived as complete

The book centers on the role of truth, more particularly Tarskian truth theory, in understanding language, thought, and communication It thus brings together issues that Davidson made prominent in philosophy over the last forty years The book divides into two parts The first is a slight revision of Davidson's Dewey lectures, "The Structure and Content of Truth," delivered in 1989 [1] The second, entirely new, centers on what Davidson calls "the problem of predication " Since Davidson sees this problem as solvable only in the context of a theory of truth, the two parts make a whole

Davidson opens chapter 1 by arguing against two types of views regarding the relation between Tarski's truth theory and philosophical issues about truth He argues against deflationary or redundancy positions that claim that Tarski's T-sentences ("'Snow is white' is true if and only if snow is white") capture all there is to the notion of truth He argues against positions that hold that Tarski's theory is irrelevant to philosophical issues because its T-sentences provide stipulative definitions of truth

[1] Published as Davidson, "The Structure and Content of Truth," this JOURNAL, LXXXVII, 6 (June 1990) 279–328

0022-362X/07/0411/580–608 © 2007 The Journal of Philosophy, Inc

Against the deflationary positions he cites quantifications over truths, which the T-sentences do not capture, and he cites the fact that Tarski's theory presupposes a nontrivial translation between object language and meta-language, which the T-sentences presuppose Against claims that Tarski's theory is irrelevant because it provides stipulative definitions, Davidson agrees with Tarski that the theory does not define the concept of truth, but captures the extensional core of an antecedently understood notion, as applied to specified languages The fact that truth has features beyond its extensional aspects (such as the fact, cited by Michael Dummett, that truth constitutes a type of success for belief and assertion) does not impugn the interest of Tarski's theory Davidson's claims here are familiar I believe them to be broadly right

Davidson uses these criticisms to motivate his construal of the truth predicate as expressing an undefined theoretical notion He takes the notion to be useful in empirical semantical theories about language Aspects of truth that Tarski's formal extensionalist theory does not make use of are relevant for understanding its place in semantical theories of individuals' speech and linguistic competence Here too Davidson's claims seem plausible They are borne out by the development of empirical semantics

In chapter 2, Davidson maintains that semantical predicates are primitives He argues against epistemic theories of truth and against what he calls correspondence theories His resistance to the view that truth is to be explained in terms of our abilities to find out whether something is true, or in terms of justified assertability, is natural and, I think, correct (33–34, 47) The views that he calls correspondence views are equally implausible Davidson counts them unintelligible On Davidson's rendering, a correspondence theory, particularly in the form of a realist view, takes truth to be "entirely independent of our beliefs" (33–34) Davidson understands correspondence views also to require that sentences or propositions have distinctive types of referents, such as facts or states of affairs (39–42, 126–27)

To be entirely independent of our beliefs, on Davidson's construal, truth would have to attach to truth bearers and falsity bearers that have natures that are independent of any of our beliefs or activities Thus meaning and propositions, true or false, would have to be what they are independently of human thought and activity Frege held such a position Russell did for a time Few other modern philosophers can be cited Although I do not see that the view is unintelligible, it is an extreme, implausible view, at least in general form The beliefs and activities of individuals play a role in determining the nature of most representational thought content, and most or all

meaning and reference in language On the other hand, I see no necessary relation between the view that Davidson rejects and correspondence theories, or even realism, as they are usually understood

Davidson somewhat obscures matters in his formulations He writes, " I was still under the influence of the idea that there is something important in the realist conception of truth—the idea that truth, and therefore reality, are (except for special cases) independent of what anyone believes or can know" (41) He refers to realism as "the doctrine that the real and true are independent of our beliefs " (42) Normally, realism centers not on the nature of truth bearers, which do plausibly depend on the thoughts and beliefs of individuals, but on truth makers A normal claim is that in many standard cases, what make truths true are individuals, properties, or what not, that do not depend for their existence or nature on our beliefs Davidson's reasons for rejecting realism and correspondence conceptions do not squarely confront this idea Much of the discussion has a coy, nonstraightforward feel to it

Davidson's construal of correspondence views as requiring distinctive types of referents for sentences also seems to me a very narrow one This construal is certainly not assumed by all who appeal to correspondence as illuminating truth Davidson's objection to correspondence views on this construal is that "there is nothing interesting or instructive to which true sentences correspond" (39, 126–27) Davidson supports this objection only by appeal to the Frege-Church-Godel slingshot argument (40, 127–30) This argument purports to show that if sentences denote anything, all true sentences denote the same thing One of the assumptions of the argument is that logically equivalent sentences correspond to (denote) the same entity Russell would certainly have rejected this assumption I believe that any philosopher who wants to maintain that true sentences correspond to relatively fine-grained entities, like structured propositions, facts, or states of affairs, would reject it I have no special brief for such views, but I find Davidson's discussion of them rather too brisk, and ultimately question begging

Tarski regarded his theory as a correspondence theory At one time Davidson agreed, maintaining that Tarski's reliance on satisfaction in systematizing truth amounts to an appeal to correspondence Davidson criticizes his former position He holds that it is contrived to see sentences as corresponding to sequences of objects that satisfy sentences And he invokes the slingshot argument (39n, 155–56) I do not find these criticisms forceful

There is much to be said for Davidson's old way of regarding correspondence One can explicate correspondence as a combination of reference and truth-of A simple subject-predicate sentence cor-

responds to the way things are just inasmuch as its terms refer to things of which its predicates are true The idea can be adapted to fit more complex sentences One need not invoke distinctive entities—*ways things are*—for *sentences* to correspond to One need not invoke a semantical *relation* beyond reference and being true of Reference and being-true-of are relations between expressions within a sentence and a further subject matter Sentences correspond to truth makers through those relations I see no harm in believing in distinctive entities like facts or states of affairs, or semantical relations between sentences and such entities But a relatively robust conception of correspondence does not depend on such a belief

I do think that it would be overblown to tout appeals to correspondence as "theories," unless one simply has Tarski-type theories in mind The defensible point in an appeal to correspondence is to insist that genuine relations between truth bearers, or their parts, and subject matters, figure in the determination of truth Such insistence is usually coupled with rejecting reductions of truth to verification, coherence, warranted assertability, redundancy, or other matters that serve to avoid giving language-subject-matter relations a role in understanding truth

Davidson would have rejected the relaxed correspondence approach that I just suggested for a reason other than those he explicitly gives He would have thought that such an approach would give explanatory roles to reference and being-true-of that they cannot bear He believed that these notions are exhaustively explainable in terms of truth I would have been interested to see more of Davidson's thinking on this issue He provides only a very brief rationale I shall return to the matter near the end of this review

In chapter 3, Davidson claims that the main aspect of the concept of truth that is left out by Tarski's extensional theory is its role as part of a framework for empirical semantical theories of individuals' utterances and linguistic competencies He holds that an applied truth theory describes a complex ability (50) He maintains that a theory of the truth conditions of sentences should be embedded in a theory of beliefs and preferences He sketches a rational reconstruction of how one could arrive at a theory of meaning and belief on the basis of knowledge of degrees to which sentences are held true—refining ideas of Frank Ramsey, Richard Jeffrey, and of course W V Quine (67–75) Davidson's sketch is very compressed But the outlines of his view of relations between truth, belief, preference, and meaning are familiar I think that taken very broadly, Davidson's sketch of the shape of empirical semantical theory is plausible and right It is one of his best contributions to philosophy

There are, however, three aspects of Davidson's vision here that I regard as mistaken or seriously blurred One is his very unclear account of whether, when, or in what way a theory of truth is internalized in ordinary speakers and hearers On one hand, he takes the theory to describe a complex ability (50) and to provide "the only way to specify the infinity of things the interpreter knows about the speaker " (52) On the other hand, he writes that the sense in which speaker and interpreter have internalized a theory "comes to no more than the fact that the speaker is able to speak as if he believed the interpreter would interpret him in the way the theory describes, and the fact that the interpreter is prepared to interpret him in this way" (53) This "as if" talk leaves some doubt about whether Davidson thought that Tarskian truth theories are psychologically real in ordinary communication Davidson's remark that language users need not have "explicit knowledge" of the theory hardly helps The question is whether, if ordinary speakers and hearers are to speak and comprehend what is said, they must have semantical concepts, concepts of utterances, unconscious beliefs in the principles, and unconscious use of the structural inferences, of semantical theories I think that despite the "as if" talk, other things that Davidson holds commit him to such a view But he does not clarify or defend his position

I think that such a commitment would be mistaken Accounting for the development of linguistic use and comprehension in children, or even in adults, appears not to require taking *semantical* theories to be tacitly realized in the individuals' psychologies [2] Attributing such tacit theories seems to hyper-intellectualize linguistic competence As far as anyone has shown, very young children can speak and understand speech without being capable of *thinking* semantically They need not have intentions about how their own speech is to be interpreted by others—the sorts of Gricean intentions that Davidson thinks are necessary for linguistic meaning (51–54) We know that most children communicate by age 14–18 months It is at best controversial whether they have concepts of belief, intention, and so on, before age three or four years A semantical theory is a conceptualized theory about semantical relations between representation and world It describes the way sentences that in fact set truth conditions relate, through their parts, to truth makers Its conceptualized semantical standpoint need not model the structure or process of individual understanding Seemingly, individuals can use sentences and their logical forms in

[2] Cf my "Comprehension and Interpretation," in Lewis Hahn, ed , *The Philosophy of Donald Davidson* (Chicago, Open Court, 1999), pp 229–50

propositional inferences without having semantical meta-concepts for understanding such uses The connections to the world could lie in nonsemantical, empirical concepts, and in unconceptualized psychological connections between perceptions of others' utterances and perceptions of coincident matters in the environment Mastery of compositionality need not be framed in semantical, meta-conceptual terms either

A second feature of Davidson's vision that I do not accept is his confining the application of semantical theory to idiolects and to what speakers fully understand I think that communal languages and dialects show enough stability for linguistic theory to apply to them, allowing of course for numerous individual variations Morever, even linguistic usage in an idiolect falls into patterns, and carries meanings, that the individual may only incompletely understand

Davidson provides ill-informed and cavalier objections to such a view He refers to those "who are pleased to hold that the meanings of words are magically independent of the speaker's intentions—for example, that they depend on how the majority or the best-informed, or the best-born, of the community in which the speaker lives speak, or perhaps how they would speak if they took enough care" (50) Davidson attributes this view to me and to Saul Kripke's Wittgenstein My view, at any rate, does not hold that meanings are independent of speaker intentions Speakers' intentions often place those speakers under standards and make them beholden to social and other mechanisms for determining reference, and thereby meaning, which they incompletely understand I have discussed these issues elsewhere, and will not press them here [3]

A third aspect of Davidson's vision that I do not accept is his account of the role of speaker-interpreter-world triangulation in the development of meaning and objectivity Davidson holds that the "ultimate source of both objectivity and communication is the triangle that, by relating speaker, interpreter, and the world, determines the contents of thought and speech" (74–75) He fully intends to maintain that before linguistic communication occurs, there is no representation or intentionality I believe that this view is incompatible with what is known about perceptual reference in a wide variety

[3] Cf my "Individualism and the Mental," *Midwest Studies in Philosophy*, IV (1979) 73–121, reprinted in *Foundations of Mind, Philosophical Essays, Volume II* (New York Oxford, 2007) with Postscript, pp 100–81, "Wherein Is Language Social?," in Alexander George ed , *Reflections on Chomsky* (New York Blackwell, 1989), pp 275–90, and "Social Anti-Individualism, Objective Reference," *Philosophy and Phenomenological Research*, LXVII (2003) 682–90, both reprinted in *Foundations of Mind*, pp 275–90 and 307–15

of animals, and about propositional inference and thought in prelinguistic children and a narrower variety of animals [4]

All three of these aspects of Davidson's position are familiar from his other writings I believe that he nowhere defends them in depth In any case, he does not advance discussion of them in this book He is concerned to sketch his view in a brief, picturesque way As indicated, I think the larger position important and plausible It is the product of insightful work, issuing from a perspective on the subject that was reflectively developed over the long term

I turn now to the previously unpublished parts of the book Davidson states "the problem of predication" in various ways He calls it the problem of explaining the semantic role of predicates (2), the problem of explaining the nature of predication (76–77), the dual problem of explaining how particulars are related to their properties and how subjects are related to predicates (77, 83) The problem is "closely related," in Davidson's view, to that of explaining the "unity of a proposition" or the unity of a judgment (4, 77) He holds that although an account that provides a solution has long existed, "no one has noticed that this account solves the problem" (77) The solution is supposed to fall directly out of Tarski's theory of truth

Davidson provides a sketchy, but interesting, historical overview of the problem He gives special attention to Plato, Aristotle, Russell, P F Strawson, Quine, Wilfrid Sellars, Frege, and Tarski The brief discussions of these figures do not purport to rest on careful historical investigation Davidson's treatment is nonetheless often deft and insightful His discussion of Plato shows, I think, a grip on the theory of Forms that bears fruit from his early work in ancient philosophy His remarks on Russell, Quine, Sellars, and Tarski are also penetrating

Davidson discusses ways in which various figures tried to formulate the problem, ways in which some fell into regress, and ways in which others avoided regress but still failed to find a solution Near the end of the book, Davidson lists four desiderata for a solution (155–59)

The first is that predication should be connected to truth

> If we can show that our account of the role of predicates is part of an explanation of the fact that sentences containing a given predicate are true or false, then we have incorporated our account of predicates into an explanation of the most obvious sense in which sentences are unified,

[4] Cf "Social Anti-Individualism, Objective Reference", and see "Perceptual Entitlement," *Philosophy and Phenomenological Research*, LXVII (2003) 503–48 I will discuss these issues at much greater length in a forthcoming book, *Origins of Objectivity*

> and so we can understand how, by using a sentence, we can make assertions and perform other speech acts (155)

I believe that this grounding of predication and propositional unity at least partly in truth constitutes an important insight The insight derives from Frege, as Davidson acknowledges I shall return to the matter of the primacy of truth among semantical concepts

Second, any attempt to provide *full* explication of the semantics of predicates by associating them with single entities of any kind (whether properties, universals, Forms, or what not) is doomed (156) The main idea is that in explaining a simple predication like 'Theaetetus sits' *simply* by postulating a semantical relation between the predicate and some repeatable entity (Sitting), one has done nothing to connect the predicate to the subject or to connect the repeatable entity to the named entity Propositional unity is not explained merely by invoking a semantical relation to a property and a semantical relation to an individual In effect, Frege makes this point decisively, and Davidson articulates it effectively throughout the book

Third, a solution must distinguish between the issue whether properties and other abstract entities exist and the issue whether predicates bear a semantical relation to such entities Davidson seems to accept that such entities exist He denies that predicates bear any semantical relation to them (158) I believe that Davidson is right to distinguish the issues, but wrong in his position on the latter one I shall return to this matter

Fourth, a solution must deal with quantificational structure As Davidson notes, Frege and Tarski advance the discussion in this respect well beyond their predecessors (158–59)

Davidson claims that Tarski's semantical method meets these desiderata and provides a solution to the problem of predication Tarski's essential innovation, in Davidson's view, is to invoke the idea that predicates are *true of* the entities named by constants or quantified through variables Davidson explains Tarski's introduction of sequences of objects The introduction enables the notion of satisfaction (roughly, the converse of truth-of) to deal with an arbitrary number of predicate argument places Davidson characterizes the circumstances under which a given sequence assigns entities to the variables as those circumstances which, if those variables were the names of those entities, would yield a true sentence Such sequences are said to *satisfy* the sentence, whether the sentence is open or closed (159, cf 36) Here, in contrast to Tarski (30, 31), Davidson explains satisfaction in terms of would-be truth (159–61, cf 36, 55n)

Davidson takes himself to have explained predication in terms of truth (161) He admits that the explanation is not fully general,

since truth is explained only for one language at a time He claims, however, that Tarski provided a method for specifying the semantical role of any given predicate Davidson believes that nothing more can reasonably be asked of a theory of predication "What more can we demand? I think the history of the subject has demonstrated that more would be less" (161)

The solution that Davidson offers is a less-is-more solution His solution avoids taking predicates to connect particulars with universals He thinks that better understanding derives from this and other sorts of avoidance I think that the conclusion of the book would have been strengthened if Davidson had discussed his conception of his solution A summary of wherein the solution does and does not connect to the issues that historically swirled around predication would have made the book's conclusion seem less abrupt

A discussion of how the solution relates to historical understandings of predication would also have made clearer exactly how Davidson sees his own role in the solution that he advocates He claims that no one had noticed that the Frege-Tarski results solved the problem He represents himself as a match-maker between solution and problem

Recall that Davidson characterizes the problem of predication as that of explaining the semantical role of predicates (2) It has long been known that the Frege-Tarski results provide a core explanation of the semantics of predication Framed this way, Davidson's claim that no one had noticed the relevance of the Frege-Tarski results to the problem of predication is mistaken

Davidson certainly contributes a thought provoking account of the history of the problem He is right to emphasize that semantical theory, centered on truth, is a good framework within which to reason about traditional ontological aspects of the problem Does Davidson himself contribute anything new to understanding predication?

I think that Davidson sees his main contribution as lying in the challenge, "What more can we demand?" (161) He believes that associating predication with the ontology of properties, universals, Forms, or the like, is a mistake He believes that many philosophers fell into regresses because they tried to explain the unity effected by predication in terms of an ontology of universals and particulars, or properties and individuals Davidson rejects such solutions I think that he believes his contribution to lie in framing problem and solution in such a way as to show that issues that traditionally were thought to be part of the problem are best separated from it altogether I am doubtful I devote the rest of this review to some of my doubts

Much of Davidson's historical discussion centers on the regress that Plato worried about in his "third-man" problem In fact, Davidson

writes, "The difficulty of avoiding one infinite regress or another might almost be said to be *the* problem of predication " (79) It is well known that Frege's approach to predication provides material to undercut regresses that are variants of Plato's third-man problem Davidson recognizes the relevance of Frege's contribution The interesting issues lie, I think, in the details regarding particular regresses

Davidson's discussions, though suggestive, are hampered by sloppy formulations that obscure premises that lead to regresses I will provide only a sampling of these discussions

The issue to be borne in mind is to what extent a regress derives from metaphysical, semantical, or syntactic commitments All discussions of the problem rightly take for granted that sentences (or thoughts) are not mere sequences of singular terms—mere lists Davidson attributes to Plato the salutary claim that every [declarative] sentence has a verb [a predicative part, at least in its logical form] This purely syntactic point can be taken as a given in any reasonable account of the semantics and metaphysics associated with sentences

Let us look at Davidson's first discussion of the regress problem

> Consider the discussion of how the forms are related to each other The question is how to think of sentences like 'Motion is not Rest' This does not deny existence to either Motion or Rest, but it does tell us that the two forms, both of which exist, are not identical, to refute Parmenides, Plato here distinguishes between the 'is' of existence and the 'is' of identity (*Sophist* 255B,C) Plato speaks of the forms as blending, connecting, or mixing with one another In the case of Rest and Motion, they fail to blend The difficulty is to reconcile these declarations with the claim that every sentence must have a verb Clearly the words 'Motion' and 'Rest' *name or refer to* forms, so if the sentence 'Motion is not Rest' has a verb, it must be 'is' or 'is not' (or 'blends with' or 'does not blend with') Plato takes Sameness and Difference to be forms, but then fails to recognize that if these forms are what is *meant* by the 'is' and 'is not' in sentences that speak of the forms blending or failing to blend, then a sentence like 'Motion is not Rest' *names* three forms ('Motion Difference Rest'?), and there is no verb
>
> If Identity and Difference are not verbs but *names of forms*, then the same must hold of 'sits' and 'flies', they should be 'Sits' and 'Flies' Both 'Theaetetus sits' and 'Motion is not Rest' lack a verb How could Plato have failed to notice this? Of course, 'Motion is not x' and 'x is not Rest' are one-place predicates However, this observation is of no help as long as one holds that predicates *refer to* forms But we have been told that every sentence must contain a verb (82–83, my italics)

If one adds a verb, the cycle starts over again Motion is related to Rest by Difference

In this passage, Davidson shifts, without comment, from "name or refer to" to "meant" and then back again [5] In the passage, he also writes of Sameness and Difference as being taken by Plato to be forms Then one sentence later, without comment, he writes of Identity and Difference as names of forms

The issues of *exactly* what semantical relation verbs or predicates bear to a subject matter, and the relative contributions of language and subject matter to predication are the central issues in understanding both predication and the regress problem It does not serve reflection to engage in careless formulations that bear on just these issues

This particular formulation of a regress argument can be rendered valid if and only if what it is for 'is not' *to mean* a Form is to be *fully explained* by what it is for 'Difference' to *name* the same Form Or more precisely, the regress gets started if and only if

'is not' means Difference

is either notionally equivalent to, or means the same as, or is interchangeable with, or is to be *fully* explained by

'Difference' names Difference

Bringing out these assumptions helps isolate what must be rejected to avoid the regress The regress gets started if the syntactic and semantical roles of a predicate are assimilated to those of a singular term It is absolutely essential to any solution that this assimilation be rejected

Inconstancy in the use of terms for semantical relations that predicates bear to entities runs through the book (86, 92, 113, 119, 134, 136, 139, 145, 158) Davidson writes as if it does not matter whether we think of predicates or verbs as naming, referring to, standing for, relating to, introducing, signifying, designating, corresponding to, adverting to single entities like properties, relations, universals, Forms, or the like These are all terms that Davidson uses

Davidson has a motive for being off-hand He thinks that no such semantical relation helps in any way to solve the problem of predication or the regress problem (85, 112, 139, 146, 156–58) That is, he thinks that taking predicates to bear *any* semantical relation to a single entity like a property, relation, universal, Form, or the like makes no contribution at all to solving these problems In fact, he takes it to be a necessary part of a solution to *reject* the idea that a

[5] Davidson does write later of *apparent* contradictions and of getting clear about the problem (83) But the formulations in the quoted passage seem to me to be needlessly unclear and inconstant

predicate bears *any* semantical relation to a single entity like a property, relation, universal, Form, or the like (85, 158) I believe that Davidson positions here are mistaken

At issue in what follows is a distinction between two types of positions One is that postulating a semantical relation between predicates and properties, relations, universals, Forms, or the like, does not *in itself* solve the problem of predication The other is that any such postulation does not *in any way* help in a solution, or even more strongly, that any such postulation should be rejected as a part of any solution

Davidson gives good reasons for the first position Just postulating a semantical relation between singular terms and single entities and a further semantical relation between predicates and single entities, like properties, does nothing in itself to explain how the semantical relation associated with the predicate connects to the semantical relation associated with the singular term to form a unity–a sentence that can be true or false (for example, 86, 92–94, 159) [6]

I think that Davidson gives no good reason for the latter views He gives no good reason to deny that predicates bear a semantical relation to properties, relations, or the like Imprecise and shifting formulations tend to camouflage the differences between the two types of views

Here is Davidson's second discussion of the regress problem

> There is no objection to taking properties and relations as entities about which we want to think and say things, unless, of course, there are no such entities I shall not cast doubt on their existence the question of whether they exist will play no part in what follows The more basic question is whether positing the existence of properties and relations *helps us understand* the structure and nature of judgments like the judgment that Theaetetus sits or that Motion is different from Rest Let us take one more look Theaetetus sits Theaetetus, we agree, is an entity, a person He is sitting The property of Sitting is another entity, this time a universal that can be instantiated by many particular entities In the present case, Theaetetus is one of those entities In other words, *we explain what it is to be the case* that Theaetetus sits by saying that Theaetetus instantiates the property of Sitting The fact itself doesn't, then, *consist*

[6] I think that Davidson articulates well the point that no semantical relation between predicates and a single entity can exhaust the semantical role of predicates This point is less ambitious than his primary thesis (that it is a mistake to invoke any such relation in predication at all) The less ambitious thesis is, I think, well known Davidson does a service, however, in insisting on it Being true of a variety of individual entities is an ineliminable element in the semantics of predication Being capable of effecting, in sentences, an application or attribution that when successful yields a true sentence, or thought, is a central aspect of a predicate's semantical role

> merely of the two entities, Theaetetus and the property of Sitting It is a fact because those two entities stand in a certain relation to each other, the relation of Instantiation Now we see that a fact we can describe in just two words apparently involves three entities But can '*Theaetetus sits*' consist of just the three entities, Theaetetus, Sitting, and Instantiation? Surely not The fact requires as well that these three entities stand in a certain relation to one another Theaetetus and Sitting, in that order, bear the relation of Instantiation to one another To explain this fact we need to mention this *fourth* entity, which, unlike Instantiation, is a three-place relation We are clearly off on an infinite regress (84–85, all italices are mine, except the last)

Davidson begins here by writing that the basic question is whether positing properties *helps us understand* the structure and nature of *judgments* Then six short sentences later he shifts to writing about our *explaining what it is to be the case* The explanation of what it is to be the case is an explanation of a *fact,* not a judgment A few sentences later, Davidson writes of the quoted *sentence* 'Theaetetus sits' as *consisting* of the individual, the property, and the relation One sentence later he shifts back to writing of what the fact "requires "

Quite apart from the error of writing of sentences as consisting of individuals, properties, and relations, there remain two important slides in this discussion One is between writing about judgments or sentences and writing about facts The other is between writing of postulating properties and relations to *help* us understand something and writing of explaining what something *is* or what it *consists in*

Suppose that there are properties and relations Given this supposition, it is not obvious that there is anything wrong with assuming that there are a lot of them In addition to the individual Theaetetus and the property Sitting, let us suppose that there is a relation of instantiation between the property and the individual, and a relation in turn among that relation, the property, and the individual, and so on to infinity Such an infinity is profusion, not circularity or regress Circularity and regress are properties of explanations—of certain representational structures—not of ordinary facts or subject matters

Davidson shifts from discussing judgments to discussing facts and what it is to be the case, then to discussing sentences, then back to discussing facts He conjures the infinity with respect to the facts and then pins the regress on the structure or nature of judgments or sentences In understanding the regress, it is crucial to distinguish between what the judgment or sentence is doing—what its semantics is—and what is the case in the world Davidson blurs these distinctions

Similarly, Davidson is very lax in his remarks about explanation His regress argument is supposed to target the idea that positing

properties *helps explain* the nature of judgment But the argument uses a premise about *explaining what something consists in* This is a difference between holding that a semantical relation between predicates and properties/relations contributes *something* to a solution and holding that it constitutes the *whole* solution

Suppose that one holds that in the judgment that Theaetetus sits, there is a predication of the property Sitting of Theaetetus Postulating a semantical relation between the predicate 'sits' and the property Sitting helps us understand, let us suppose, the semantical structure and nature of the judgment It is not the full story about the predication The semantical relation between predicate and property is only an aspect of predication Let us say that the predicate *indicates* (or predicatively denotes) the property or relation Beyond indication, predication also constitutively involves a capacity for *application* or *attribution* of the property to the individual Thus predication involves both indication and application of what is indicated to something further

It would be a mistake to claim that the judgment consists in a predication of a further relation (Instantiation) between individual and property The judgment does not predicate the relation of instantiation to the individual and the property Perhaps, necessarily, Theaetetus sits if and only if Theaetetus instantiates the property Sitting But a judgment that Theaetetus sits is *a different judgment, involving a different predication,* from a judgment that Theaetetus instantiates the property Sitting A semantical explanation of the latter judgment can legitimately be different from a semantical explanation of the former

A semantical explanation is not an exhaustive definitional or explanatory substitute, or a revelation of what a judgment "consists in " The paradoxes teach us that much The point is, I think, intuitively clear even before reflecting on the paradoxes Semantical explanations help us understand the semantical structure of judgments and sentences They are not definitional substitutes or exhaustive explanatory replacements for what they help us understand

So far Davidson's account gives no reason to deny a semantical relation (I have called it *indication* or predicative denotation) between predicates, on one hand, and properties or relations, on the other The relation is not name-like No name can enter into the very same relation Indication essentially involves a potential for predication Predication essentially involves (at least purported) application or attribution as well as (at least purported) indication (cf note 16) Finally, a predication of the property Sitting is not the same as, nor can it be exhaustively and definitionally explained by, a predication of the relation Instantiation between the property Sitting and an individual We can learn these things from regress arguments

Davidson continues the passage just quoted by trying to simplify the regress argument

> The problem is easier to state in semantical terms The sentence 'Theaetetus sits' has a word that refers to, or names, Theaetetus, and a word whose function *is somehow explained* by mentioning the property (or form or universal) of Sitting But the sentence *says* that Theaetetus *has* this property If the semantics of the sentence were *exhausted* by referring to the two entities Theaetetus and the property of Sitting, it would be just a string of names, we would ask where the verb was The verb, we understand, *expresses* the relation of instantiation Our policy, however, is to *explain verbs by somehow relating them* to properties and relations But this cannot be the end of the matter, since we now have three entities, a person, a property, and a relation, but no verb When we supply the appropriate verb, we will be forced to the next step, and so on (85–86, my italics, except for the italics on "has")

There are two errors in this argument One is the claim that 'Theaetetus sits' *says* that Theaetetus has the property of Sitting The original sentence does not say that Theaetetus has the property Sitting If it did, the sentence would predicate the relation Having of the two named entities, Theaetetus and Sitting The sentence does not do that In fact, we may suppose, it only predicates the property Sitting of Theaetetus If Theaetetus sits, Theaetetus has the property Sitting But the predications in the two sentences (or judgments) are different The syntactical and semantical structures of the sentences are different[7] Claiming that 'Theaetetus sits' says that Theaetetus instantiates or has the property Sitting amounts to assimilating 'sits' to a singular term and then looking elsewhere for a verb Assimilating the semantic role of a predicate to that of a singular term is the cardinal error in falling into regress Davidson does not mark this claim separately as the cardinal error, because he thinks that the fundamental error lies further upstream He thinks that the root mistake lies in taking the sentence's predicate to "mention" the property Sitting He does not seem to realize that assimilating the semantical role of a predicate to that of a singular term (in fact, the cardinal error) does not follow from the assumption that predicates bear a semantical relation to properties like Sitting

The second error in this simplified regress argument is the move from the idea (a) that an explanation of predication "*mentions*" (bears some semantical relation to) a property to the idea (b) that the verb *expresses* the *relation* between the property and the individual to which

[7] Davidson makes the same mistake on p 147

it is attributed Davidson seems to sanction this error It is a *non-sequitur* This second error is closely related to the first It matters *exactly* what the semantical relation is between an element of a sentence or judgment and an element in the world If predicates are to be semantically related to properties (if they are to indicate properties), they must also attribute the properties They do *not* predicate, or express *in the same semantical relation*, a further relation between the properties and entities that they attribute the properties to (cf note 17)

Davidson accuses many thinkers of falling into Platonic third-man regresses, or failing to confront the problem of predication at all I find his discussions of Hume, Bradley, Russell, Quine, and Sellars better than those of Kant,[8] Strawson,[9] or Frege However, most of the discussions, even those that are penetrating, move too quickly to provide solid grounding for substantive claims

Since I think that Frege made the deepest contribution to understanding predication, I will discuss Davidson's engagement with him Davidson recognizes Frege's contribution to be pivotal He highlights these contributions as well as weaknesses in Frege's positions

Frege sharply distinguishes the grammatical positions of predicates and singular terms He associates the grammatical distinction with a difference in semantical role He holds that singular expressions denote objects He thinks that predicates denote functions When a (first-level) predicate is embedded in a sentence and the singular-

[8] Davidson writes that Kant said little that was relevant to the problem of predication (99) I think that this claim is mistaken Kant has a serious and extensive theory of the unity of judgment Some indication of it occurs at *Critique of Pure Reason*, B140–42 Stripped of its connection to idealism, I think that Kant's account comprises major contributions in its associating predication with judgment, and its connecting predication to a system, to objectivity, and to truth—roughly, "objective validity " For Kant, the system is the unity of apperception I leave open here whether this is the right system to relate truth to Frege made do with a system of inferential connections without insisting that the system must involve self-consciousness In some of these respects, Frege's breakthroughs go back to Kant Of course, Kant did not exploit them remotely as deeply as Frege did

[9] Regarding Strawson, Davidson writes, "It is interesting that someone who made it "central" to his life's work to explain 'singular reference together with predication' should, so far as I know, have paid no serious attention to what I am calling the problem of predication" (113) I think that Davidson's sarcastic criticism of Strawson for ignoring the problem of predication involves serious misrepresentation For example, he attributes to Strawson the view that the "entire semantic function" of 'pretty' in 'Sally is pretty' is to designate a general or universal entity (113) Davidson quotes nothing from Strawson to support this attribution I have not been able to find anything in Strawson's article "Singular Terms and Predication" in P F Strawson, *Logico-Linguistic Papers* (London, UK. Methuen, 1971), to which Davidson refers, to support it Although I think that Strawson does not provide a clear, strong account of these matters, his discussion of grouping seems to me to count against Davidson's attribution

term positions in the predicate are filled by singular terms instead of quantifiable variables, the singular terms denote arguments or inputs for the functions Given an input, the functions yield truth or falsity as values, or output The importance of this account is multiform

First, it provides a sharp distinction between both grammatical and semantical roles for singular terms and predicates Predicates cannot operate, syntactically or semantically, as names or singular terms Singular terms cannot operate, syntactically or semantically, as predicates Although Frege calls the semantical relations that singular terms and predicates enter into "denotation" (*Bedeutung*), the relations are different, as his logical grammar makes clear Let us call the first relation *singular-denotation* and the second *predicative-denotation* The semantics of predication is ineliminably that of predicative application or attribution For Frege, predicative application is application of a function Predicative application occurs when predicates are embedded in a sentence When a predicate is joined with a term, the predicate's predicatively-denoted function is given an argument or input to which the function is represented as applied The application yields a value or output An essential aspect of the semantical role of predicates is their potential for predicative application or attribution I think it best to see Frege's contribution as transcending the particular type of entity, a function, that he took predicates to apply when they are joined with terms One could give up Frege's invocation of functions as predicative denotations of predicates, and regard predicates as applying properties, relations, universals, or themselves, to something else In doing so successfully, they are true of the something else, and they characterize it

Second, the account sets both singular-denotation and predicative-denotation in essential relation to the truth or falsity of sentences or thoughts Frege points out that the concern with language-world and thought-world relations is inseparable from our concern with truth value

Third, Frege provides a model for understanding predicative application that illuminates the grammatical unity of sentences and thoughts in terms of a semantical unification Modeling predication on functional application indicates how the truth value of a whole sentence, or thought, functionally depends on and only on the denotations applied by the predicates and the denotations of the singular terms (granted the particular order and form of those applications)

Fourth, Frege extends the account to quantification in a smooth and systematic way

Davidson praises the second and third contributions thus "Of all the efforts to account for the role of predicates that we have reviewed,

Frege's is the only one that, by its treatment of predicates, clearly makes sentences semantic units Of the attempts we have considered, Frege alone has assigned a semantic role to predicates which promises to explain how sentences are connected to truth values" (133–34) Davidson also gives Frege credit, though I think less fully, for his deep fourth contribution (147, 158–59)

Davidson does not accept Frege's view that predicates literally denote functions from entities to truth-values He thinks, and I agree, that truth values are not plausibly regarded as objects It should be said that Frege's view is still invoked in logic and linguistics for its elegance and convenience

Although it is natural and probably correct to deny that sentences denote truth values as objects, I think that such denial is not of *deep* conceptual importance Davidson thinks otherwise He follows Dummett in holding that this denial targets a very serious mistake on Frege's part Davidson endorses Dummett's intemperate characterization of Frege's position as a "gratuitous blunder" (136) Davidson follows Dummett in charging that Frege made the mistake of assimilating sentences to names I have shown that Dummett's view rests on an elementary misunderstanding of Frege's logical theory [10] Although Frege holds that sentences and names bear the same denotation-relation to certain objects, he firmly distinguishes names and sentences in other ways, even within his formal theory For most purposes, it is a minor issue whether sentences bear a semantical relation to truth values, regarded as objects If one avoids this view, one can retain Frege's insight that the semantics of predicates combines with the semantics of other types of expressions *functionally* to determine the truth or falsity of sentences Application of a predicate's denotation to a singular term's denotation *uniquely* determines the truth or falsity of a sentence that consists of such a predication Similarly, for sentences of more complex syntactical types Frege's insight stands, regardless of whether predicates are taken as literally applying functions, and regardless of whether truth values are taken as objects

The question is whether in giving up the idea that predicates literally denote functions whose values are truth values, one must give up the idea that they bear a semantical relation to properties, relations, or the like Davidson believes that one must I think that he gives no good grounds for his belief

[10] Cf my "Frege on Truth," in L Haaparanta and J Hintikka, eds , *Frege Synthesized* (Boston Reidel, 1986), pp 97–154, section II, reprinted in *Truth, Thought, Reason Essays on Frege* (New York Oxford, 2005), cf pp 108–15 For a shorter account, see the Introduction to *Truth, Thought, Reason*, pp 21–23

To motivate his belief, Davidson discusses a notorious problem for Frege. Frege held that singular terms cannot singularly-denote functions, and functional expressions cannot predicatively-denote objects. Frege realized that this claim entails that 'the concept *horse*' cannot singularly-denote a concept. (Concepts for Frege are functions whose values are truth values.) He concluded that the concept *horse* is not a concept. Of course, this conclusion is counterintuitive. I think it mistaken. Eventually Frege rejected, for science, the predicate 'is a concept', which he thought could be true only of objects, not concepts.

Davidson's assessment of these matters differs from mine. Here is Davidson's account:

> One [serious problem] was pointed out by Frege: the entities referred to by predicates cannot be objects, like the entities referred to by singular terms, for if they were, sentences would be strings of names, and the usual problem of how a sentence could constitute a unit would once more emerge. Frege therefore stressed the contrast between the referents of singular terms ('objects') and the referents of predicates ('concepts'). Yet here, Frege admits, we face a dilemma, for as soon as we say anything about a concept we convert it into an object. (134)

Davidson believes that these difficulties should be jettisoned simply by rejecting the view that predicates bear any semantical relation to concepts, functions, properties, relations, or the like.

It is unclear *exactly* why Frege believed that the entities predicatively-denoted by predicates cannot be singularly-denoted by singular terms. Whatever Frege's reason, he was mistaken. I will return to this point. First, let us consider Davidson's reasons, on Frege's behalf, for holding that entities predicatively-denoted cannot be singularly-denoted.

Davidson asserts that if the entities "referred to" by predicates could be the entities "referred to" by singular terms, sentences would be strings of names. This claim is correct, but it is misleading. It is true that if predicates bore the same semantical relation (reference) to the same types of entities that singular terms do, they would *be* singular terms, and sentences would be lists. Davidson's reasoning is misleading on two counts. First, Frege did not take predicates to bear the same semantical relation to functions that singular terms bear to their referents. Second, the key to understanding why sentences are not strings of names lies not in distinguishing the entities that names and predicates bear semantical relations to. The key lies in distinguishing the relevant semantical relations themselves. Frege draws this latter distinction, but Davidson's reasoning assimilates predicate "reference" to singular term "reference."

For Frege predicates and singular terms bear different semantical relations to their subject matters. Frege seems not to have recognized

clearly that his distinction between these semantical relations does not force him into his mistaken view that singular terms cannot singularly-denote what predicates predicatively-denote So let me argue the point independently of Frege's full position There is a difference in the semantical relations that predicates and singular terms bear to the entities that they are semantically related to, *whether or not the entities that they bear semantical relations to are disjoint* (cf note 12) The semantical relation that predicates enter into is not merely that of being in a semantical relation, predicative denotation, to functions The semantical relation for predicates also constitutively involves a capacity for predicative application, or attribution When their argument places are filled by singular terms, predicates also connect functions to arguments That is, one aspect of the semantical role of predicates is to effect a representational analog of functional application Singular terms never have this semantical role Even if they were to singularly-denote functions, singular terms would be unable to represent application of the functions to inputs into the functions Since this capacity is an aspect of the semantical relation between predicates and functions, singular terms cannot bear the same semantical relation to functions that predicates do, regardless of whether singular terms can bear a semantical relation to functions Singular terms lack the potential for representing functional application when they are "connected" with singular terms Connecting one singular term with another produces a list It does not yield a representation of functional application Predicative-denotation is different in this respect So singular-denotation and predicative-denotation must be distinguished And this distinction does not entail that singular terms cannot singularly-denote what predicates predicatively-denote

Frege distinguishes the two relations By writing of the two semantical relations indiscriminately as a "reference" relation, on the mere supposition that both relations relate expressions to single entities, Davidson conflates the two relations in a way that Frege does not [11]

Suppose, against Frege, that singular terms can singularly-denote functions It does not follow that sentences become mere lists The difference between predication and singular reference can be pre-

[11] For more on these matters, see Montgomery Furth, "Two Types of Denotation" in *Studies in Logical Theory*, American Philosophical Quarterly Monograph Series 2 (Oxford, UK. Blackwell, 1968) As I shall explain, I believe that although Frege distinguished the two types of denotation, he failed to realize that this distinction obviated any need to require that singular terms could not bear *their* semantical relation to the same entities that predicates bear their different semantical relation to (namely, to functions)

served To distinguish sentences from lists, one must maintain a difference in semantical role between singular terms and predicates, and one must maintain that a capacity for application be counted part of the semantical role of predicates Frege thought of this role literally as function-application I think it better not to hold that predicates literally predicatively-denote functions The applicational aspect of the semantical role of predicates is better regarded as attribution (of properties or relations) that *functionally determines* the truth or falsity of sentences or thoughts One need not take truth values as objects Whether the difference between singular terms and predicates lies in *entities* that singular terms and predicates are semantically related to is not crucial What is crucial is that there be a difference in the semantical relations, and that the semantical relation for predicates include a potential for application or attribution [12] So the reason that Davidson gives for agreeing with Frege that any entities that predicates are semantically related to cannot be referred to by singular terms is not a good one His talk of "conversion" of functions into objects is unsupported metaphor

Davidson is right to maintain that a relation between predicates and entities such as properties, relations, functions, universals, and so on, cannot exhaust the semantical role of predicates If it did, one could not explain the unity of sentences or thoughts One would leave out predicative application—attribution But Davidson's reasoning depends on the idea that predicate "reference" exhausts the semantical role of predicates Frege did not hold this view The view does not follow from the mere claim that predicates bear some semantical relation to single entities such as properties, relations, or functions I have called this relation "indication" or "predicative denotation "

Davidson gives no good reasons to support Frege's claim that singular terms cannot singularly-denote what predicates predicatively-denote (namely, functions) Frege's claim was mistaken I believe that Davidson's reasoning shows that he did not sufficiently appreciate the importance of the first of Frege's contributions to understanding predication that I listed above Frege's sharp distinction between the semantical roles of predicates and singular terms and his semantical association of predication with functional application blocks regress The distinction allows predicates to bear a semantical relation to

[12] I do think that certain entities are not suited to being predicated For example, physical individuals are not I think that sets are not Aristotle's view that there are certain metaphysical limitations on what can be a predicable seems to me to be correct But I believe that *any* entity can in principle be singularly-denoted or named Predication and singular reference do not yield a disjoint metaphysical division of the world

single entities (such as functions or properties) other than individuals of which the predicates are true [13]

I believe that Frege himself did not fully appreciate the importance of his contribution He thought that a difference in the semantical relations that predicates and singular terms bear to their subject matters has to correspond to a strict division in the subject matters that they could be related to In fact, the sharp distinction between semantical relations, one of which constitutively involves a potential for predicative application, is sufficient to account for the semantical distinctiveness of predication, without invoking a metaphysical divide between entities that predicates and singular terms can be semantically related to This is what Frege seems not to have realized

Whatever Frege's reasons for holding that singular terms cannot singularly-denote any function that a predicate predicatively-denotes, they were mistaken Church showed in his calculus of lambda conversion that there is no incoherence in taking functional expressions, including predicates, that contain open argument places to bear a semantical relation to the same entities that nonfunctional expressions, which lack open argument places, bear a (different) semantical relation to [14] It does not follow from holding that predicates predicatively-denote properties, relations, or functions, that there is anything wrong with allowing singular terms to singularly-denote the same entities What is critical is that the functional expressions also have the role of representing application of those entities to other entities, when the functional expressions are grammatically joined with singular terms in sentences or thoughts To avoid trouble in a Fregean theory, it is enough that one distinguish the semantical relations and maintain certain grammatical type differences

A primary theme in my discussion has been that Davidson thinks that a successful account must hold that predicates bear no semantical relation to single entities like properties, relations, universals,

[13] Frege himself was not open to such a view I will discuss his mistake further on in the text below Frege sometimes characterizes an object as anything that is not a function, and says that an expression that denotes an object is a name or singular term Cf Gottlob Frege, *Kleine Schriften*, I Angellelli, ed (Hildesheim, Germany Georg Olms, 1967), p 134, p 18 in the original, *Die Grundlagen der Arithmetik*, translated as *The Foundations of Arithmetic*, J L Austin, trans (Oxford, UK. Blackwell, 1980), pp x, 77n, 72) Even regarded as definitions, these characterizations are, I think, mistaken Frege shows that he has an independent grip on both the grammatical category of name or singular term and on the notion of an object I discuss these matters in some depth in "Frege on Truth," in *Truth, Thought, Reason*, pp 101ff There I also discuss Frege's mistaken doctrine that singular terms cannot singularly-denote functions

[14] Alonzo Church, *The Calculi of Lambda-Conversion* (Princeton University Press, 1941) For further discussion see my *Truth, Thought, Reason*, pp 20–21

Forms, or functions (82, 84–86, 93–95, 97, 112, 114, 134–35, 139, 143, 145–46, 156, 158, 161) Some of his arguments depend on careless formulation (82, 84–86, 143, 145–46) Some depend on not distinguishing clearly between showing that any such semantical relation cannot *exhaust* the semantical role of predicates (they must be capable of functional application or attribution) and arguments that such a semantical relation *cannot figure at all* in predication (114, 156, 158) None of the arguments is sound

I want now to remark on Davidson's positive position, laying aside his arguments for it Davidson sensibly does not question the existence of properties or relations Such entities are very commonly referred to and quantified over in scientific explanations Nominalist or extensionalist rejection of their existence seems to me to be a pointless and unscientific position In this respect, Davidson differs from Quine and Sellars His ontological liberalism always seemed to me to be a modern and enlightened aspect of his philosophical work However, he agrees with Quine and Sellars in rejecting the idea that such entities figure in the semantics of predication Davidson thinks that the only way to bear a semantical relation to them is through singular reference, or quantification into the position of singular reference [15]

I believe that this position is very unattractive If properties are admitted to exist, it is completely implausible to claim that predicates bear *no* systematic semantical relation to them—including no relation that bears systematically on the truth of sentences containing such predicates A sufficient condition for a semantical relation to be associated with a subsentential linguistic expression is roughly that there be a systematic relation between the expression and an entity in a subject matter, a relation that affects the truth value of sentences containing the expression, that is induced by the use of the expression,

[15] As Parsons remarks in a footnote that Cavell includes (146n5), Davidson does accept a semantical relation between predicates and *certain* "single entities"—sets He does not confine himself strictly to the semantical relation of being true of Davidson's view here is a concession to Tarski's method of explaining satisfaction as a relation between predicates and sequences of objects It should be said that Davidson firmly and correctly rejects the idea that the semantical relation of predicates to sets or sequences exhausts their semantical role (157–58) As Parsons notes, however, it is unclear how Davidson might defend his allowance of a semantical relation between predicates and sets, given his arguments that a semantical relation between predicates and "single entities" like properties, relations, and Forms, are a source of regress All that Davidson says in favor of preferring sets to properties is that their individuation conditions are clearer than properties (119)—a familiar, and I think unpersuasive, point from Quine—and that sets and not properties are appealed to in Tarski's truth theory Neither point seems relevant to showing that regress looms with properties but not with sets

and that is in principle understandable by understanding the language [16] The combination of ideas (a) that there is a property Sitting and (b) that the verb 'sits' bears *no* systematic semantical relation to that property when 'sits' is used in sentences is, I think, on its face completely implausible

The combination would amount to denying that there is a systematic relation, understandable by understanding the language, between 'sits' and Sitting For if 'Sitting' names a property, and 'sits' is conceptually and systematically related to 'Sitting', then 'sits' cannot avoid bearing some *semantical* relation to the property The relation differs from that of singular reference But it is a systematic, linguistically understandable relation that is induced by the normal use of the language and that bears on the truth value of containing sentences Understanding this relation need not go through understanding the relation between 'sits' and 'Sitting' An individual could have failed to learn the gerund construction, and used only 'sit' The idea that the individual's word 'sit' does not bear a systematic semantical relation to the property Sitting, granted that the property exists, is, I think, untenable

Of course, 'sits' bears systematic relations, constitutively associated with its use, to other entities too It bears systematic relations to the

[16] This rough sufficient condition is too narrow to function as a necessary and sufficient condition It leaves out not only semantical relations for thoughts It also leaves out semantical relations between linguistic expressions and meanings, concepts, senses that are ways of representing a subject matter, rather than constituting a (first-level) subject matter The rough sufficient condition will suffice for present purposes

I should say here that a full account of predication must include an account of the representational content of predication (the mode of predication), not simply an account of what is predicated through predicative denotation and of predicative application or attribution Davidson's account centers entirely on the subject-matter aspects of predication I believe that one needs a notion of concept, or predicative representational content, to account for the way predications are thought The existence of nonlinguistic thought is only one of several obstacles to resting content with the predicates themselves Systematic explanation calls for abstraction that comprises representational content in addition to linguistic items and what they represent I mention two familiar reasons for understanding the semantics of predication in this two-level way There are different ways of thinking of any given property or relation in predications of that property or relation Such differences emerge in psychological and psycholinguistic explanations Further, in some cases, there is no property or relation that is attributed or applied, yet there can remain a predicative thought I leave open the relation between concepts and linguistic meanings I believe that certain aspects of meaning, over and above reference, should be distinguished from what individuals think with (namely, their concepts) in carrying out predications in thought But this is a complex issue that would take us far afield Since Davidson focuses on the simplest issues regarding the relation between predication and truth, I believe that his omission of discussion of further aspects of predication does not seriously harm his project Davidson has a more Quinean view of these matters than I do, but these differences between his view and mine are not forced to the surface by the issues Davidson discusses in this book

individuals that sit, to the set of individuals that sit, to a concept, or way of thinking, that one learns when one learns such words as 'sits' and 'sitzt' And so on Understanding predication and how language works in relation to the world, and acknowledging the existence of these entities, requires recognizing that there are systematic semantical relations between 'sit' and whatever among these entities exist (I think that all of them, including properties and relations, do) Some may be less central to a particular style of explanation than others But ontological liberalism seems to me to require liberalism regarding semantical relations that words bear to the world [17]

What is important for present purposes is not whether properties and relations exist The key point is that no dire difficulty in understanding *predication* arises if one assumes that predicates bear a semantical relation to them

Sometimes Davidson claims that properties are not *needed* for semantical explanation (112, 139, 158) Tarski's truth theory does not appeal to them The fact that Tarski truth theory makes do with sets and not properties does not show that a semantical theory that invokes properties is of no explanatory worth Appeal to a semantical relation between predicates and properties cannot *suffice* to explain predication But it is natural and so far undefeated to think that predicates attribute properties to entities denoted by singular terms

[17] It should be noted that this sort of semantical liberalism will allow not only a semantical relation between 'sits' and Sitting, but also a different semantical relation between 'sits' and Instantiation of Sitting, and perhaps a further semantical relation between 'sits' and Instantiation of Instantiation of Sitting, and so on If there is a profusion of relations in the subject matter, there is nothing wrong with there being a profusion of semantical relations to elements of the profusion that can be understood by understanding the language To see that no regress is involved, it is important to note the following facts All such semantical relations are different semantical relations 'Sits' only *predicates* Sitting, not all of the other entities Predication with 'Sits' is apriori associable with those other entities But it is not to be explained in terms of them Rather, the reverse is true By understanding predication, and accepting the ontology, one is in a position to understand each of the infinitely many further semantical relations The statements of these further semantical relations are entailments of the statement containing the simple predication But these statements have different propositional contents, and understanding them is not necessary for understanding the original predication (except insofar as understanding entailments always deepens understanding) Competence to understand the whole infinity can be generated from understanding what is involved in the semantics of the original predication, and accepting the ontology One need not understand the infinity to understand the particular predication In fact, I think one can understand the particular predication itself without invoking any semantical notions, including the notion *predication*, at all What semantical notions help us understand are semantical relations I think that it is clear that the semantical relations governing predication are more basic in explaining the semantical workings of language and thought than the further semantical relations involving instantiations

One of Davidson's insights, articulated in the first half of the book, is that Tarski's theory provides a *limited* account of truth The theory bears only on *extensional* aspects of truth It does not illuminate connections to belief, preference, and meaning that Davidson rightly believes are part of applications of the theory Insofar as truth is essential to understanding predication, Tarski's theory must also be seen as a limited theory of predication In being a purely extensionalist theory of truth, the theory yields only an extensionalist theory of predication It omits aspects of predication that constitutively relate to belief, preference, meaning, and even to singular reference (to properties and relations), in common-sense and scientific theorizing So the argument that Tarski's theory does not appeal to properties, relations, or the like in its account of predication does not support a denial that a full account of predication should invoke semantical relations to such entities

Full understanding of predication cannot rest with the semantics of predication in *sentences* in any case Since some thought is language-independent, a theory must explain the role of predication in thought The concerns of Kant and Russell to understand the mind's role in unifying judgments and propositions are relevant to full understanding Davidson has little to say about mental agency in predication This omission is perhaps partly the product of his belief that thought does not precede language and is not in any way independent of it Even if that belief were true, however, the structural issues that Davidson concentrates upon do not exhaust the subject The role of mental agency in predication would remain worthy of discussion

I want now to return to a matter that I raised earlier I wrote that Davidson would have been dissatisfied with explicating correspondence in terms of reference and being-true-of He believes that reference and being-true-of are not semantical relations with *any* independent explanatory force Davidson purports to explain predication and satisfaction (hence being true of) *fully* in terms of truth (159, 161, 163) He also tries to explain singular reference fully in terms of truth (159–61, 36, 55n) For him, there is only one semantical primitive, the concept of truth The semantical notions for singular reference and predication are derivative This view contrasts with the view that truth, predication, and reference are reciprocally inter-dependent primitive notions, and that in different respects each illuminates the others [18,19]

[18] Davidson's view also contrasts with the view that truth is fully explainable in terms of reference and/or being true of In contrast to the view that maintains that the semantical notions are reciprocally related, I do not take this sort of view seriously

[19] Davidson's actual explanation of satisfaction (159), in terms of the truth of would-be sentences that result from substituting names for variables, fails to confront well-

Davidson gives only an unelaborated reason for characterizing reference and being-true-of as, in effect, whatever relations suffice to enable a truth theory to deduce the intuitively correct T-sentences He holds that the only testable implications of a truth theory are the T-sentences Since the T-sentences do not contain the notions of reference or being true of, he claims that any way of characterizing reference and satisfaction that "yields confirmable T-sentences is as good as another" (55n) He thinks that we have no other grip on these other semantical notions than as *some* semantical relations between the parts of sentences and a subject matter that are necessary to derive the T-sentences It is only the T-sentences that we understand and can confirm in empirical semantical theories So the semantical relations between parts of sentences and subject matter are to be explained as whatever relations make the T-sentences come out true

This argument seems to rely on a mistaken view of scientific theories The idea that we explain theoretical relations as whatever make our tests come out right, and that the only constraint on a theory is that it derive testable consequences, is, I think, clearly unacceptable

A more interesting mistake underlies the argument The argument is symptomatic of a tendency in the philosophy of language to ignore the way empirical linguistic reference and linguistic predication are grounded in perception [20] We have an independent source for understanding both of these notions through their connection to singular and attributive elements in pre-linguistic perception Thus I believe that these semantical notions cannot be exhaustively explained in terms of their contribution to the derivation of T-sentences We often assess the reference of a name or demonstrative construction in terms of what we know about *what* an individual perceives And we assess the attribution of many predicates in terms of what we know about *how* an individual perceives what he perceives—how he perceives particulars *as being*

Of course, singular and predicative elements in sentences or thoughts that rely on perception are inextricably associated with propositional beliefs and propositional inferences In seeking to understand such

---

known problems about nondenumerable subject matters I think that Davidson's explanation does not provide an adequate account of sentences involving quantification over nondenumerable subject matters

[20] Examples of this sort of thinking also occur in W V Quine, "Ontological Relativity" in *Ontological Relativity and Other Essays* (New York Columbia, 1969), pp 26–68, Hilary Putnam, "Models and Reality" (1978) in *Realism and Reason, Philosophical Papers,* Volume III (New York Cambridge, 1983), pp 1–25, Putnam, *Reason, Truth, and History* (New York Cambridge, 1981), chapter 2 I find all of these ways of thinking quite wayward

elements, we necessarily integrate our understanding of them into a structure of propositional inference and belief Belief and propositional inference are necessarily understood in terms of truth conditions Truth is involved in the constitutive representational function of belief and inference Reference and predication have their point largely in their relation to truth So the notions of singular reference and predication in sentences and thoughts must be understood in relation to the notion of truth So these notions are not independent of the notion of truth This was Frege's second insight

On the other hand, our notion of truth is partly explicated in terms of notions of reference and predication, which constitutively involves the notion of being true of These notions are constrained by applications of our notion of perception (which helps ground the basic sort of belief, *de re* belief) The notion of perception is independent of any notion of propositional truth [21] So our notion of propositional truth cannot exhaustively explain our notions of reference and predication Reference and predication help explain how truth is related to truth makers

Our understanding of predication is enriched not only by what we know about perception It is enriched by all that we know about properties and relations We understand predication in connection to those aspects of a subject matter that go beyond the minimum needed to engage in extensional derivations of theorems of a Tarskian truth theory The three semantical notions are primitive Truth bears *reciprocal* constitutive relations to singular reference and predication

Davidson's book offers many insights It also has serious defects Exposition and argumentation, particularly in the second half of the book, leave much to be desired Although the work is stimulating in its discussion of predication and historical treatments of it, and clearly the product of substantial philosophical talent, I believe that it makes no significant new contribution to the understanding of predication On the other hand, Davidson *has* made significant contributions to the understanding of truth Utilizing the formal and empirical work of Frege and the formal work of Tarski, Davidson framed a broad and lasting account of the form of *empirical* semantical theory

[21] Perception is constitutively associated with conditions for veridicality But accuracy in perception is not propositional truth There is substantial evidence that animals that lack propositional attitudes, such as bees, have visual perceptual systems, capable of perceiving and perceptually attributing physical entities such as sizes, shapes, colors, distances, and so on There is substantial evidence that these same animals lack any propositional attitudes Moreover, propositional structure is not necessary for explaining the structural organization of perceptual representational content So perception bears no constitutive relation to any states that are evaluable for propositional truth

before any one else did so He also provided examples of empirical theory application that helped get empirical semantics started [22] The first half of this book sets his account in the context of others' thought about truth In this context, it continues to look strong Sad, that there will be no more work from this creative source

TYLER BURGE

University of California, at Los Angeles

[22] Cf for example, Davidson's "The Logical Form of Action Sentences" (1967), reprinted in *Essays on Actions and Events* (New York Oxford, 1981), pp 105–22

CAMBRIDGE

New and Noteworthy Titles

Aristotle's Ethics as First Philosophy
Claudia Baracchi
$90 00 Hb 978 0-521 86658-3 352 pp

Morality and Political Violence
C A J Coady
$80 00 Hb 978 0-521-56000-9 332 pp
$27 99 Pb 978 0 521-70548 6

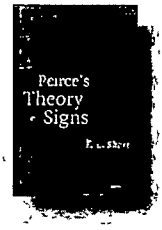
The Cambridge Companion to Renaissance Philosophy
Edited by James Hankins
Cambridge Companions to Philosophy
$95 00 Hb 978-0-521-84648 6 448 pp
$29 99 Pb 978 0 521 60893-0

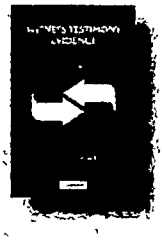
Now in paperback!
Mainstream and Formal Epistemology
Vincent Hendricks
$23 99 Pb 978 0-521-71898-1 208 pp

Debating Design
From Darwin to DNA
Edited by William A Dembski and Michael Ruse
$50 00 Hb 978-0-521 82949 6 424 pp
$24 99 Pb 978 0 521-70990-3

Coming soon in paperback!
Darwinism and its Discontents
Michael Ruse
$19 99 Pb 978 0-521-72824 9 275 pp

Prices subject to change





The Cambridge Companion to the Philosophy of Biology
Edited by David L Hull and Michael Ruse
Cambridge Companions to Philosophy
$90 00 Hb 978 0-521-85128 2 552 pp
$34 99 Pb 978 0-521-61671-3

The Limits of Loyalty
Simon Keller
$85 00 Hb 978 0-521 87461 8 246 pp

Medieval Jewish Philosophical Writings
Edited by Charles Manekin
Cambridge Texts in the History of Philosophy
$90 00 Hb 978 0-521 84023-1 304 pp
$29 99 Pb 978 0 521-54951 6

Peirce's Theory of Signs
T L Short
$91 00 Hb 978 0 521 84320 1 392 pp

Body Consciousness
A Philosophy of Mindfulness and Somaesthetics
Richard Shusterman
$85 00 Hb 978-0-521-85890 8 240 pp
$24 99 Pb 978 0 521 67587-1

Witness Testimony Evidence
Argumentation and the Law
Douglas Walton
$85 00 Hb 978-0-521-88143-2 384 pp
$29 99 Pb 978 0 521 70770-1

Aristotle on Homonymy
Dialectic and Science
Julie K Ward
$85 00 Hb 978-0 521 87486 1 230 pp

www cambridge org/us/philosophy

CAMBRIDGE UNIVERSITY PRESS

NEW TITLES IN PHILOSOPHY FROM CORNELL UNIVERSITY PRESS

FREEDOM IN THE WORKPLACE?
GERTRUDE EZORSKY

"This clearly written and tightly argued text introduces students to moral concepts of freedom, force, and coercion In addition to offering an original account of these important moral concepts, it teaches students about working conditions for the bottom tier of American workers Ezorsky illustrates her points with real world cases that vividly illustrate the stakes behind our moral concepts"

—Elizabeth Anderson, University of Michigan

104 PAGES, $12 95 PAPER/$39 95 CLOTH

ANABASIS OF CYRUS
XENOPHON

TRANSLATED BY WAYNE AMBLER
INTRODUCTION BY ERIC BUZZETTI

In this new translation, Ambler achieves a masterful combination of liveliness and a fidelity to the original uncommon in other versions, highlighting both Xenophon's gripping narrative and his sophisticated political philosophy

280 PAGES, $16 95 PAPER/$49 95 CLOTH
AGORA EDITIONS

New in Paperback

THE CONSTITUTION OF SELVES
MARYA SCHECHTMAN

"This excellent and engaging book succeeds in raising questions about the dominant approach to asking questions about our identities and our concern for the future, as well as in offering the beginnings of an alternative way to ask and answer such questions"

—*Ethics*

192 PAGES, $19 95 PAPER

New in Paperback

AQUINAS'S MORAL THEORY
Essays in Honor of Norman Kretzmann

EDITED BY SCOTT MACDONALD AND ELEONORE STUMP

"This book is historically and philosophically solid and intellectually stimulating It is an important work on Aquinas's moral thought"

—*Philosophical Review*

296 PAGES, $29 95 PAPER

NEW TITLES IN THE *ANCIENT COMMENTATORS ON ARISTOTLE* SERIES

ON PROVIDENCE
PROCLUS

TRANSLATED BY CARLOS STEEL

136 PAGES | $79 95 CLOTH

ON PLATO'S *CRATYLUS*
PROCLUS

TRANSLATED BY BRIAN DUVICK

224 PAGES | $79 95 CLOTH

www.cornellpress.cornell.edu

New books from The MIT Press

The Really Hard Problem

Meaning in a Material World

Owen Flanagan

"With an open mind, good humor, encyclopedic knowledge, and philosophical tenacity, Flanagan tackles the Big Question Can we find Meaning and Truth at the same time? Great reading for Homo sapiens " — Joshua Greene, Department of Psychology, Harvard University

304 pp , 1 illus $27 95 cloth

Subjectivity and Otherness

A Philosophical Reading of Lacan

Lorenzo Chiesa

"This is a dynamic, unsurpassed reading of Lacan that will ignite new interest in his work and rekindle the passion of initiates " — Joan Copjec, author of *Imagine There's No Woman*

Short Circuits series

368 pp , 23 illus $19 95 paper

The Chomsky Effect

A Radical Works Beyond the Ivory Tower

Robert F Barsky

Noam Chomsky as political gadfly, groundbreaking scholar, and intellectual guru key issues in Chomsky's career and the sometimes contentious reception to his ideas

416 pp , 11 illus $29 95 cloth

Ingmar Bergman, Cinematic Philosopher

Reflections on His Creativity

Irving Singer

"The work of a man who has long lived with and often written about film, this concise volume on Bergman says more than many another longwinded and less stimulating one " — John Simon, Critic

240 pp $24 95 cloth

Describing Inner Experience?

Proponent Meets Skeptic

Russell T Hurlburt and Eric Schwitzgebel

"Where many works in consciousness studies gesture at cross-disciplinary appeal, the meeting in this book of psychologist and philosopher on specific common ground puts this promise into practice " — John Sutton, Macquarie University

416 pp $34 cloth

Gödel, Putnam, and Functionalism

A New Reading of Representation and Reality

Jeff Buechner

"A must read for anyone interested in the debate Hilary Putnam started on the role, if any, of theorems in mathematical logic and computer science in computational functionalism "

— Alan Berger, Brandeis University, and Director of the Kripke Center at CUNY Graduate Center

364 pp , 1 illus $38 cloth

now in paperback

The Evolution of Morality

Richard Joyce

"Joyce's book is brilliant There is nothing more important than knowing what we are doing when we speak in the language of value We are animals that judge with cognitive and affective equipment Joyce explains who we are Nothing matters more "

— Owen Flanagan, Duke University

Life and Mind series • A Bradford Book

288 pp , 4 illus $18 paper

To order call **800-405-1619**

http //mitpress mit edu

New from OXFORD

Winner of the Lakatos Prize in Philosophy of Science, 2006

INVENTING TEMPERATURE

Measurement and Scientific Progress

HASOK CHANG

"A wonderful synthesis of the history and philosophy of physics It combines rich historical detail with philosophical acuity and imagination"—Jeremy Butterfield, Oxford University

(Oxford Studies in the Philosophy of Science)
2007 304 pp paper $24 95

WHAT ARE WE?

A Study in Personal Ontology

ERIC T OLSON

What Are We? is the first general study of personal identity in a metaphysical context It explains the title question's meaning, considers in detail the main possible answers to it, and suggests how the problem might be solved

(Philosophy of Mind Series)
2007 264 pp $39 95

LANGUAGE IN CONTEXT

Selected Essays

JASON STANLEY

Clear and concise throughout, *Language in Context* will be of interest to philosophers, linguists, and cognitive scientists who work on language acquisition, production, and interpretation

2007 256 pp cloth $80 00 paper $30 00

REVELATION

From Metaphor to Analogy

Second Edition

RICHARD SWINBURNE

"Swinburne has once again produced an outstanding work in philosophy of religion"—*The Philosophical Review*

This entirely rewritten second edition contains a long new chapter examining whether traditional Christian claims about personal morality can be regarded as revealed truths

2007 376 pp cloth $99 00 paper $40 00

MORAL UNDERSTANDINGS

A Feminist Study in Ethics

Second Edition

MARGARET URBAN WALKER

This second, revised edition contains a new preface, a substantive postscript to Chapter One, a new chapter on the importance of emotion in practices of responsibility, and a new afterword, which responds to critics of the book

(Studies in Feminist Philosophy)
2007 328 pp cloth $99 00 paper $29 95

THE NATURE OF NORMATIVITY

RALPH WEDGWOOD

The Nature of Normativity presents a complete theory about the nature of normative thought — that is, the sort of thought that is concerned with what ought to be the case, or what we ought to do or think

2007 296 pp $50 00

Prices are subject to change and apply only in the US
To order please call 1-800-451-7556
In Canada call 1-800-387-8020

Never miss an Oxford sale!
Visit our web site at www oup com/us

OXFORD UNIVERSITY PRESS

CENTRAL LIBRARY
Calcutta
18.7.08
UNIVERSITY
OF
CALCUTTA

THE JOURNAL OF PHILOSOPHY

VOLUME CIV, NUMBER 12
DECEMBER 2007

page

609 *The Two Worlds of Deflationism* — Gurpreet Rattan
639 *Three Kinds of Idealization* — Michael Weisberg
660 INDEX TO VOLUME CIV

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Wolfgang Mann, Mary Mothersill, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C Varzi *Consulting Editors* David Albert, James T Higginbotham, Robert May, Charles D Parsons, Wilfried Sieg *Managing Editor* John Smylie

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President*, Arthur C Danto, *Vice President*, Akeel Bilgrami, *Secretary*, Michael J Mooney, *Treasurer*, Liaquat Ahamed, *Other Trustees* Lee Bollinger, Leigh S Cauman, Barbara Gimbel, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

You may also visit our website at www journalofphilosophy org

THE JOURNAL OF PHILOSOPHY

2008

SUBSCRIPTIONS (12 issues)

Individuals	$45 00
Libraries and Institutions	$100 00
Libraries and Institutions (print and online)	$180 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for advertising rates, ad space is limited, so ad reservations are required

Published monthly as of January 1977, typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027 (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U S copyright law All rights reserved For more information about our copyright policies, please consult our January issue or contact us

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc , MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027 Periodicals postage paid at New York, NY, and other mailing offices

POSTMASTER. Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027

© Copyright 2007 by the Journal of Philosophy, Inc

ISSN 0022–362X

CENTRAL LIBRARY
UNIVERSITY OF CALCUTTA
KORE CAMPUS

THE JOURNAL OF PHILOSOPHY
VOLUME CIV, NO 12, DECEMBER 2007

THE TWO WORLDS OF DEFLATIONISM*

Here is how Paul Horwich describes the philosophy behind deflationism (what he calls "minimalism") about truth

> One may or not propose an account which inextricably links truth with other matters for example, assertion, verification, reference, meaning, success, or logical entailment Minimalism involves the contention that truth has a certain purity—that our understanding of it is independent of other ideas [1]

The view that our understanding of truth is independent of these other ideas leaves open the question of upon what our understanding *is* dependent The deflationary view is that the concept of truth is completely exhausted by something like that the entailment of which Tarski used to constrain an adequate definition of truth—either the *disquotational schema*, or the *equivalence schema*, or, to use the term associated with Tarski, the *T-schema* At this general level, deflationism consists of the following two tenets that our understanding of truth

(1) involves the (nonparadoxical instances of) the T-schema, and

(2) does not involve anything besides this, in particular it does not involve connections between truth and other ideas

(1) expresses the positive account of our understanding of truth, (2) expresses what is deflationary about the view

I want to argue that this philosophy is incoherent The argument proceeds in three stages

* This paper has benefited from the critical feedback of Ray Buchanan, John Collins, Matti Eklund, Ray Jackendoff, Philip Kremer, Douglas Patterson, Stephen Schiffer, James Woodbridge, and Cory D Wright, as well as by the comments of anonymous referees Thanks to them all

[1] Horwich, *Truth* (New York Oxford, 1990), p 12

0022-362X/07/0412/609–38 © 2007 The Journal of Philosophy, Inc

The first stage consists of a discussion of what might seem a distant topic of a semantically excessive interpretation of certain phenomena of interest to cognitive science (henceforth *the semantically excessive interpretation*) The semantic excess of the semantically excessive interpretation consists in the *two-worlds structure* it imposes on the understanding of representation (section II) The lesson for semantics is that there is to be a fundamental separation between the theory of representation and the theory of truth I argue that this separation has grave philosophical consequences, pushing the semantically excessive interpretation to what I call quasi-technically "the worst possible theory of truth" (section III)

The semantically excessive interpretation ceases to look like a distant topic when it is recognized that the philosophy behind deflationism *also* insists on a separation of our understanding of truth from—what is already prominent in Horwich's list—ideas from the theory of representation The second stage of the argument explains and defends this connection between the views with the aim of showing, through the same considerations and argument, that deflationism, too, is committed to the worst possible theory of truth (section IV) I also consider and reply to some objections to this stage of argument (section V)

Being committed to the worst possible theory of truth is not the worst thing in the world But things get worse for deflationism In the third and concluding stage of argument, I note two differences between the semantically excessive interpretation and deflationism, and argue that those differences further entrench the case against deflationism The semantically excessive interpretation is *merely* committed to the worst possible theory of truth Deflationism is, in addition, subject to a charge of internal inconsistency or incoherence Its positive commitment to understanding truth in terms of the T-schema cannot be combined with its deflationary commitment to purity and independence from other ideas (section VI)

I begin with some introductory and absolutely basic remarks on the concept of truth These remarks follow up, and expand upon, the positive deflationist commitment embodied in (1) The philosophical annotations that I append to the positive commitment of deflationism—embodied in the commitment to (1)—already contain the incoherence in deflationism, but I leave the final unpacking of the charge of incoherence to section IV

I A PLATITUDE ABOUT TRUTH

It should be agreed that any plausible account of truth begins with, even if it does not end with, something like Alfred Tarski's criterion of

adequacy for a definition of truth for a language [2] According to Tarski, an adequate theory of truth for a language L will satisfy *Convention T* by implying, for each declarative sentence of the language as consequences of the definition, an instance of the T-schema

(T) *s* is true-in-L if and only if *p*,

instances of which are

(TI) 'snow is white' is true-in-English iff snow is white

and

(TIF) 'la neige est blanche' is true-in-French iff snow is white

The T-schema is meant to provide at least a partial *characterization* of the concept truth—an elucidation of our understanding of truth It is worth being explicit how it is intended to do so, and from where the plausibility that it *can* do so is supposed to derive

Consider again (TI) (and (TIF)) The mentioned sentence on the left-hand side (LHS) means something in English, namely that snow is white The sentence on the right-hand side (RHS), in virtue of being a used version of the very same sentence (or a translation of that sentence), is a way of saying what the mentioned sentence says, namely, that snow is white Now, if things are as a use of the mentioned sentence (or a translation of it), says they are, that is, if snow is white, then the mentioned sentence is true, and, if things are not as a use of the mentioned sentence (or a translation of it), says they are, that is, if snow is not white, the mentioned sentence is false In other words,

(TI) 'snow is white' is true-in-English iff snow is white

What we have done here is derived an instance of the T-schema from a more general, underlying, indeed platitudinous idea the idea, namely, that a mentioned sentence is true iff things are as a use of it (or a translation of it), says they are, in this case, iff, snow is white [3]

The T-schema characterizes the concept of truth because the instances that comprise it conform to this platitudinous thought that

[2] Tarski, "The Concept of Truth in Formalized Languages," in *Logic, Semantics, Metamathematics*, J H Woodger, trans (New York Oxford, 1936), pp 152–278, and "The Semantic Conception of Truth," *Philosophy and Phenomenological Research*, IV (1944) 341–75

[3] For related discussion, cf §3 of John McDowell, "Anti-Realism and the Epistemology of Understanding" (1981), and §2 of his "In Defence of Modesty" (1987), both reprinted in his *Meaning, Knowledge, and Reality* (Cambridge Harvard, 1998), pp 314–44 and 87–108, respectively Cf also the introduction of Crispin Wright, *Realism, Meaning, and Truth* (Cambridge Blackwell, 1993, 2nd ed), pp 1–43, at p 18, and the more opaque

connects the concepts of truth and meaning or what a sentence says We can put the platitude underlying the T-schema in another, slightly more formal way namely, with

(Plat) if s means-in-L that p, then s is true-in-L iff p

Since instances of the T-schema are such that the used sentence $\lceil p \rceil$ is constrained to mean what the mentioned sentence $\lceil s \rceil$ means, the instances of the T-schema conform to the platitude The force of the idea that an adequate definition of truth should entail the instances of (T) comes from the fact that the instances of (T), like (TI) and (TIF), conform to and, when deployed in a characterizing role for the concept of truth, express a commitment to this fundamental platitude

These points are of general significance and bearing, and hold despite differences at lower levels of generality More specifically, these points need not be substantially affected either by differences in what are taken to be the bearers of truth—propositions, thoughts, mental attitudes, utterances, and so on—or by different accounts of the logical form of truth ascriptions Consider first the issues of the bearers of truth For bearers of truth different from sentences, platitudes analogous to (Plat) can be constructed to connect the truth of the bearer to what the bearer represents For example, in the case of propositions, we have (TP)

(TP) the proposition that p is true iff p,

with an instance of (TP) being

(TPI) the proposition that snow is white is true iff snow is white

The underlying platitude here is just

(Plat*) if the proposition that p represents (the state of affairs) that p, then the proposition that p is true iff p [4]

---

formulation in Donald Davidson, "Truth and Meaning" (1967), reprinted in his *Inquiries into Truth and Interpretation* (New York Oxford, 1984), pp 17–43, at pp 22–23 My formulations (like those to be found in the passages from McDowell, Wright, and Davidson that I have just cited) ignore the complications introduced by context-dependency

[4] Hartry Field writes "on most conceptions of proposition, the question of what it is for a proposition to be true is of little interest" (Field, "Critical Notice Paul Horwich's *Truth*," *Philosophy of Science*, LIX (1992) 321–30), because propositions are, on these conceptions, individuated by their truth conditions Field goes on to suggest that what is wanted by the propositional deflationist is a notion of proposition that can rekindle the interest in deflationism by making it a controversial doctrine, and this, Field thinks, will send the propositional deflationist back to a concern for, broadly speaking, linguistic deflationism (Field prefers utterance and mental attitude defla-

Perhaps similar (although perhaps more complicated) things can be said for the other candidate truth bearers If that is correct, it does no harm to stick to the linguistic version, and I do so, for ease of exposition, among other reasons [5]

Consider, second, a view that involves a different account of the logical form of truth ascriptions I have presupposed that truth ascriptions take the form of a predication of truth to bearers, but 'is true' may be construed, for example, as a prosentence forming operator that produces special types of sentences that are dependent on anaphoric connections to secure their content Deflationism can and has been put in this way, with analogues (differing in logical form) of (T) and (TI) functioning to characterize the concept of truth [6] But they can function this way only because these analogues to (T) and (TI) conform to something like (Plat) The prosentence formed by appending 'is true' to a referring term *t* that refers, say, to the sentence 'snow is white', inherits its content anaphorically from the content of the sentence to which *t* refers, that is, from the content of 'snow is white' So the prosentential analogue to (TI) will be such that the sentence on which the prosentence forming operator 'is true' operates will have the same content as the sentence used on the RHS of a prosentential analogue of (TI), namely the content that snow is white So the prosentential analogue to (TI) will conform to (Plat) [7]

---

tionism) My proposal is different but similar in spirit It also relies on *not* thinking of propositions as, by definition, individuated by their truth conditions, but instead of thinking of them, by definition, as representing things essentially representing what they do The connection between representation and truth conditions in (Plat*) then makes the connection between *propositions* and truth conditions that traditional accounts of propositions take as definitional I do not claim that this sounds very natural But I do want to suggest that (TP), (TPI), and (Plat*) can rekindle the interest in propositional deflationism If that is right, then it can be said that (Plat) and (Plat*) ground the truth of (TI) and (TPI), which in turn play identical theoretical roles for the different kinds of deflationists—once one has chosen sides about what the primary bearers of truth are, (TI) and (TPI) will partially define what it is for a bearer to be true The focus here is on these roles assigned by deflationism *despite* the connections made explicit in (Plat) and (Plat*)

[5] The focus on sentences seems especially appropriate in the discussion of the semantically excessive interpretation given the background assumption that thought is carried out in some symbolic system of representation Also, this is an appropriate place to make clear that I am using the expressions 'meaning', 'content', and 'representation' (when it does not mean representational vehicle) pretty much interchangeably

[6] See Dorothy Grover, Joseph Camp, and Nuel Belnap, "A Prosentential Theory of Truth," *Philosophical Studies*, XXVIII (1975) 73–125

[7] Cf Douglas Patterson, "Theories of Truth and Convention-T," *Philosopher's Imprint*, II (2002) 1–16, http //www philosophersimprint org/002005/, at p 6, and Michael Williams, "Meaning and Deflationary Truth," this JOURNAL, XCVI, 11 (November 1999) 545–64, at p 546

So the argument of this paper is intended to operate at a high level of generality, in which the philosophy behind deflationism is challenged in the light of the role of the family of platitudes of which (Plat) is our representative member The arguments of sections III and IV trace out the consequences with which this family of platitudes about representation and truth is pregnant But before moving on to that, I introduce and discuss the semantically excessive interpretation

II THE TWO-WORLDS STRUCTURE IN COGNITIVE SCIENCE (AND BEYOND)

I want to discuss a natural, influential, and in some inchoate way, perhaps even intuitive, interpretation of certain phenomena of interest to cognitive science I will argue that the interpretation is semantically excessive in its understanding of representation In motivating the semantically excessive interpretation, I will use an argument reconstructed from Ray Jackendoff's[8] critique of the "naive (and nearly universally accepted) [idea] that the information language conveys is about the real world "[9] The semantically excessive interpretation is both philosophically familiar, and familiar from the work of other linguists, psychologists, and cognitive scientists, but I think that Jackendoff's view is especially interesting in its particular, and quite explicit, blending of psychological, linguistic, and metaphysical considerations [10]

According to Jackendoff, phenomena of interest in the cognitive science of perception reveal the extent to which the general principles of our visual system shape, or structure, or somehow enter, the content of perceptual experience For example, a consideration of the Rubin Vase/Face illusion, in which the imposition of figure and ground relations allows one to see a two-dimensional figure as a vase with a black background or as two faces with a white background, but not both, shows that

> what one sees cannot be solely environmental in origin, since the figures are imbued with organization that is not there in any physical sense, that is by no means logically necessary, and [which may be] inconsistent with the physical facts [11]

[8] Jackendoff, *Semantics and Cognition* (Cambridge MIT, 1983), "The Problem of Reality," *Noûs*, XXV (1991) 411–33, *Foundations of Language* (New York Oxford, 2002)

[9] *Semantics and Cognition*, p 24

[10] I discuss, below, some cognate philosophical views For related views from linguistics, see, for example, Noam Chomsky, *Knowledge of Language Its Nature, Origins, and Use* (New York Praeger, 1986), at p 45, and Emmon Bach, *Informal Lectures on Formal Semantics* (Albany SUNY, 1989), at p 98 Jackendoff's *Semantics and Cognition* account strikes me as the most radical and uncompromising statement of the position

[11] *Semantics and Cognition*, p 25

Jackendoff supports the point through a consideration of related visual phenomena, such as other perceptual illusions (like the Hering-Helmhotz illusion in which parallel lines appear differently bowed when contrasted against "concave" and "convex" patterns of straight lines), and perceptual ambiguity, in which a figure can be seen as, say, a duck or a rabbit, but not both at once [12] In each case, Jackendoff's contention is that there is a shaping of the content of perceptual experience that comes from the mind and not from reality Call this thesis, applied generally to cognitive systems, the *Principle-Ladenness of Content*

These issues from the cognitive science of perception lead Jackendoff to his first point that the cognitive principles of the perceptual system shape the content of perceptual experience But the main point is not confined to the content of perceptual experience Jackendoff contends as well that these kinds of considerations have repercussions for thought and natural language The connection to thought and natural language are made through Jackendoff's *Cognitive Constraint*

> There must be levels of mental representation at which information conveyed by language is compatible with information from other peripheral systems such as vision, nonverbal audition, smell, kinesthesia, and so forth If there were no such levels, it would be impossible to use language to report sensory input [13]

The Cognitive Constraint requires compatibility across linguistic, thought, and peripheral system contents, so as to allow language to be able to be used to report the contents of peripheral systems One might find the idea of a compatibility of contents across systems attractive without thinking that that compatibility is demanded by facts about reportability I think that is right, but for our purposes, it is only the compatibility, and not its precise basis, that matters

One bold way to meet the Cognitive Constraint is to simply identify the kind of information conveyed by language with the kind of information conveyed by the peripheral systems Jackendoff calls this identification the *Conceptual Structure Hypothesis*

> There is a single level of mental representation, conceptual structure, at which linguistic, sensory, and motor information are compatible (*ibid*, p 17)

This view makes compatible linguistic, thought, and peripheral systems content by invoking conceptual structure as a shared level of mental representation [14]

[12] See Ludwig Wittgenstein, *Philosophical Investigations* (Oxford Blackwell, 1953), Part II, xi

[13] *Semantics and Cognition*, p 16

[14] Jackendoff intends a *syntactic* characterization of the idea of a shared level of mental representation, what is required is compatibility in kinds mental representa-

Although controversial, none of this seems theoretically excessive or heavy-handed But Jackendoff takes the claim that the cognitive principles of the visual system shape the content of perceptual experience to contain radical consequences for our understanding of representation In particular, the role the cognitive principles of the visual system play in shaping the content of perceptual experience prevents the content of perceptual experience from being about *the real world*, rather, the content of perceptual experience is about a world that is the product of a causal interaction between the real world and the cognitive principles of the visual (and other) systems—it is about what Jackendoff calls *the projected world* This claim, together with the Conceptual Structure Hypothesis, entails that language, thought, and perception carry content only *about* the projected world Jackendoff explicitly draws the conclusion

> It should now be clear why we must take issue with the naive position that the information conveyed by language is about the real world We have conscious access only to the projected world—the world as unconsciously organized by the mind, and we can talk about things only insofar as they have achieved mental representation through these processes of organization Hence *the information conveyed by language must be about the projected world* [15]

So Jackendoff's argument runs as follows, with P3 or *Semantic Excess* being the objectionable premise

(P1) The cognitive principles of systems shape the contents of the representations used in those systems For example, the cognitive principles of the visual system shape the content of the perceptual representations tokened in perceptual experience (Principle-Ladenness of Content)

---

tions construed as pieces of mental syntax But Jackendoff also employs a notion of the information carried by a representation, of what the representation is about I think that this is also the case in later work, despite a disavowal of representations and representation talk in favor of the idea of "cognitive structures," and how they work by "making distinctions" (*Foundations of Language*, p 20) One needs, I think, some notion of content here (one that does not prejudge the issue of whether the content is about reality or the "reality" that is constructed by our perceptual systems—see the following note) For related discussion, centered on Chomsky's disavowal of representations and representation talk in theorizing linguistic competence, see Georges Rey, "Chomsky, Intentionality, and a CRTT", and Chomsky, "Reply to Rey," both in L Antony and N Hornstein eds , *Chomsky and His Critics* (Malden, MA Blackwell, 2003), pp 105–39 and 274–87, respectively

[15] *Semantics and Cognition*, p 29, emphasis in original This understanding of representation survives into Jackendoff's later work, where representation is again concerned with "the 'reality' that is constructed by our perceptual systems in response to whatever is 'really out there' *the perceptual world is reality for us*" (*Foundations of Language*, p 308, emphasis in original)

(P2) Compatibility amongst linguistic, thought, and peripheral system contents is required (Cognitive Constraint), assume this compatibility to consist in the fact that linguistic, thought, and peripheral system contents involve a shared level of mental representation, *conceptual structure* (Conceptual Structure Hypothesis)

(P3) If the cognitive principles of a system shape the content of representations used by that system, then those contents are of a projected world rather than of the real world (Semantic Excess)

(C) The contents of the representations used in linguistic, thought, and peripheral systems are of a projected world rather than of the real world

Our concern will be to understand and evaluate Semantic Excess

Simplifying, but not too much, Jackendoff's understanding of representation is as follows

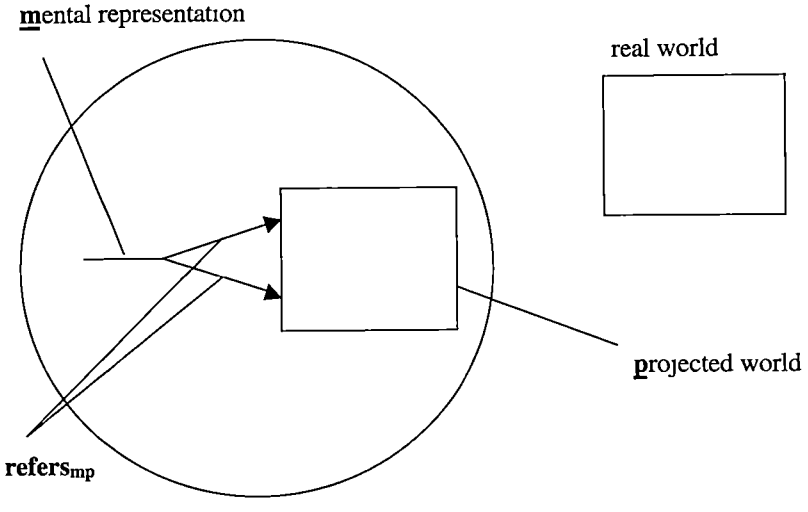


Figure 1 Representation and reality in Jackendoff's cognitive science

Figure 1 highlights *two absolutely crucial commitments* of the view The commitments are also crucial to the arguments to follow against the semantically excessive view and against deflationism, in the sense that the applicability of the arguments requires an understanding of the views as involving precisely these commitments (I discuss views to which the argument does not apply below) The first commitment is the following Jackendoff has no intention of falling into an idealism according to which *all that there is* are the entities of the projected world ('#entities#' in Jackendoff's terminology—more on this in a moment) He is not taking what he calls a "solipsist position", it is not

being claimed "there is no real world", rather, it is being claimed "only that [the real world] is not what we see "[16] This is reflected in Figure 1, which shows the real world as a separate realm [17] Second, and, crucially, *therefore*, we cannot, construe our representations as being semantically related to the real world, but only to the projected world This is reflected in Figure 1 in the reference relation More generally, semantic relations do not hold between representations and the real world, they hold, instead, between representations and the projected world

What this shows is that "we must question the centrality to natural language semantics of the notions of truth and reference as traditionally conceived "[18] In understanding this we must keep Jackendoff's anti-idealism firmly in mind Jackendoff is *not* arguing that the traditional notion of truth, for example, *is* a semantic notion but that it is misunderstood, and that the entire world is a projection His point, rather, is that there is a traditional notion of truth, of how things are in the real world, and that this traditional notion does not play a role in semantic theory The lesson for semantics is that it is truth-in-the-projection, and not truth, that is the central notion of a semantic theory This separation of representation from truth, and its attachment to some notion explicitly contrasted with truth, is the imposition of the two-worlds structure on our understanding of representation

In any event, the charge of semantic excess is directed against views that have, and substantially exploit, this kind of two-worlds structure in understanding representation

I said that Jackendoff's view is interesting in the way that it blends a wide range of considerations, but it is also interesting in how obviously it parallels some fundamental problems in the history of philosophy Insofar as there are genuine parallels here, the argument of this paper—which is directed against deflationism—can be extended to and run off of one of the historical cases I consider briefly below I will also indicate the kinds of view to which the argument does not apply

A more comprehensive understanding and argument against deflationism would see it, I think, against the background of a problematic history of two-worlds responses to skepticism In our example, the

[16] *Semantics and Cognition*, p 29

[17] Although there are causal relations that obtain between the real world and an individual's body and sensory mechanisms, these relations are left unspecified in Figure 1 because they are not *semantically* relevant—they have nothing to do constitutively on Jackendoff's view, as far as I can see, with the representational properties of mind and language Cf Jackendoff, *The Architecture of the Language Faculty* (Cambridge MIT, 1997), chapter 8 (especially p 191)

[18] *Semantics and Cognition*, p 29

skepticism lies in the background of, and motivates, Semantic Excess, for if the cognitive principles of the visual system shape the content of perceptual experience, 'how can perceptual experience furnish us with knowledge of a mind-independent world?' The two-world structure response is also clear it cannot We have access only to the projected world, and our claims are semantically related to this projected world This rehearses the kind of dialectic that leads to an interpretation of Kant's transcendental idealism that is two-worlds in structure, with our representations making reference to the phenomenal *world* rather than to the noumenal *world* Both the two-worlds structure, as well as the skepticism to which it is a response, are at least in broad outline philosophically familiar I contend that this two-worlds interpretation of transcendental idealism is subject to an argument exactly parallel to that which is to follow

A subtler and less well-known case involves a change in the logical positivist theory of meaning—the theory according to which, in the slogan, meaning is method of verification The change concerns the understanding of verification as *conclusive verification*, which is supposed to be a factive notion, to an understanding of verification as *confirmation*, which is not factive The move from factive conclusive verification to nonfactive confirmation, where confirmation is *explicitly contrasted* with conclusive verification, together with the idea that meaning is method of verification, again sets up the two-worlds structure [19] In this case, I think the presence of the two-worlds structure is a response, ultimately, to skeptical problems posed by a factualist or cognitive understanding of philosophical discourse and disagreement The view is subject to an argument exactly parallel to that which is to follow

The argument to follow does not, however, apply to more radical forms of idealism The view, does not apply, for example, to a Berkeleyan idealism in which the world represented is the world as conceived in the mind of God The argument will also not apply to contemporary views which demand the "unboundedness of the conceptual"[20] or demand that the "world is the world as it figures in our world view,"[21] or that construe the central notion of a theory of representation as "truth as conceived by the agent "[22] These views are,

[19] For this kind of move, see Rudolf Carnap, "Testability and Meaning," *Philosophy of Science*, III (1936) 419–71, §1

[20] McDowell, *Mind and World* (Cambridge Harvard, 1994), Lecture II

[21] McDowell, "Rehabilitating Objectivity," in Robert Brandom, ed , *Rorty and His Critics* (Malden, MA Blackwell, 2000), pp 109–23, at p 115

[22] Akeel Bilgrami, *Belief and Meaning* (Cambridge Blackwell, 1992), for example, at pp 139–40

or seem, idealist in character, but they do not impose the two-worlds structure on the understanding of representation From the point of view of the argument to be given here, these views are on par with a naïve realism, in which thought is straightforwardly and unproblematically about the world

Again, the argument to be given here does not apply to a view of the nature of semantics that one *can* cull from Chomsky and some of those who follow him For example, Paul Pietroski argues that an understanding of semantics according to which semantics is truth conditional is inconsistent with an understanding of semantics according to which semantic values are determined compositionally [23] Pietroski argues that semantic values should not be construed as real-world objects and truth values (or truth conditions), but rather as *whatever values* can figure in a compositional account of the language The key difference between this view and Jackendoff's is that *speakers* are still able to refer to real-world objects, and to say that things that are true or false, it is just that *words* themselves do not have semantic values so as to be able to do this Again there is no two-worlds structure what *a speaker* says is about, ultimately the one and only world that there is

To take another example, Peter Ludlow argues that semantics and metaphysics can mutually benefit by seeing through to the possibility of a referential semantics for Chomskyan I-languages [24] I-languages are in the first instance not conceived referentially or truth conditionally, but instead conceived roughly as cognitive principles of the language faculty (like the cognitive principles of the visual faculty) A referential semantics for I-languages incorporates I-substances—substances to which we can refer only because we have an I-language of a certain kind—as referents But it may also incorporate P-substances—substances that are referred to but to which reference can be made without having an I-language of a certain kind This view does not involve detaching the reference relation for any expression from P-substances, and re-attaching them to I-substance correlates, with truth being a matter of how things are with the no-longer-referred-to P-substance and I-truth being a matter of how things are with the referred to I-substance correlate On Ludlow's view, different referents have different ontological statuses, but the truth of a representation is just a matter of how things are with referents, whatever their individual ontological statuses This view

[23] Pietroski, "The Character of Natural Language Semantics," in Alex Barber, ed , *Epistemology of Language* (New York Oxford, 2003), pp 217–56

[24] Ludlow, "Referential Semantics for I-Languages," in Hornstein and Antony, eds , pp 140–61, §5

does not separate representation from truth So there is no two-worlds structure in this view

Let me summarize Jackendoff offers an interpretation of certain cognitive scientific phenomena that is two-worlds in structure, and in which the fundamental semantic notions of reference and truth are to be replaced by their -in-the-projection counterparts The problem with the view, from the point of view of the argument to be given here, is not that that the interpretation has idealist tendencies, nor even that it offers an alternative to truth conditional semantics It is the *specific way* that it expresses its idealist tendencies, and it is the *kind of* alternative to truth conditional semantics that it offers The problem is that the semantically excessive view separates representation and truth to produce a two-worlds structure in the understanding of representation This view has grave philosophical consequences

III SEMANTIC EXCESS

I want now to explain and defend the charge of semantic excess against the semantically excessive interpretation

Tarski notes, in considering an objection to his semantic conception of truth, that

> [w]e may accept the semantic conception of truth without giving up any epistemological attitude that we may have had, we may remain naïve realists, critical realists or idealists, empiricists, or metaphysicians—whatever we were before The semantic conception of truth is completely neutral toward all these issues [25]

One way to read Tarski here is as saying that the semantic conception of truth is meant to be compatible with a wide range of further commitments about truth Many different theories of truth are compatible with the requirement that a theory of truth entail instances of the T-schema Perhaps it is too much to ask that every theory of truth *entail* the instances of the T-schema [26] It will, nevertheless, still be requirement on any theory of truth that it be capable of recovering an interpretation of the instances of the T-schema in such a way that they come out as true, and obviously so If this is right, any theory of truth that cannot interpret the instances of the T-schema as true, and obviously so, will be a very bad theory, in fact, such theories are so bad that I will say that they commit themselves to *the worst possible theory of truth* The worst possible theory

[25] "The Semantic Conception of Truth," p 362
[26] See Patterson, §2

of truth is that equivalence class of theories that cannot interpret the instances of the T-schema as true, and obviously so

What can Jackendoff say about the instances of the T-schema? Well a quick argument shows that he cannot interpret them in such a way that they are true and obviously so Remember the *two absolutely crucial* commitments of Jackendoff's view, as I am interpreting it to distinguish between the notions of truth and truth-in-the-projection, and to press for the semantic relevance of the latter as opposed to the former (cf Figure 1) But in that case, the use of a mentioned sentence does not specify a *truth* condition for a sentence, but rather a *truth-in-the-projection* condition for a sentence So, for example, although it may be true, it is not obviously true, that

(TI) 'snow is white' is true iff snow is white[27]

but only that

(TIJ) 'snow is white' is true-in-the-projection iff snow is white

So it looks like the theory is committed, in our technical sense, to the worst possible theory of truth Of course, the idea behind my obviously tendentious terminology—'the worst possible theory of truth'—is that a failure to interpret the instances of (T) as true and obviously so, renders a theory absurd as a theory of *truth* I will elaborate and defend this quick argument in a moment, but first I want to introduce one more element of Jackendoff's position

It is extremely revealing that Jackendoff introduces a special notation for referring to elements of the projected world reference is made to elements of the projected world by enclosing expressions in #-marks, for example, '#snow#' refers to snow in the projection (or, better snow-in-the-projection) But Jackendoff, as theorist, also wishes to make reference to elements of the real world, and for that the words are simply used, unadorned, without any distinguishing marks #-less expressions can be used in the language of the theorist, then, to refer to the real world Jackendoff makes the following remarks about his proposal

> The procedure may at first seem paradoxical having just denied the possibility of referring to the real world, I am nonetheless introducing a metalanguage that purports to do so To be clear about what is going on, however, we must keep in mind that, like most theories, the present theory is stated from the point of view of a hypothetical omniscient

[27] To keep things manageable, from here on in, I suppress the reference to English in the truth predicate

> observer The theory claims that [speakers] can speak only of the projected world—but it can only do so only by assuming tacitly that we, as hypothetically omniscient observers, are not so limited [28]

So the hypothetically omniscient can refer, in thought and language, to the real world, but real thinkers and speakers can refer only to the projected world

Significantly, Jackendoff allows that for some expressions, both #-ed and #-less versions refer, for example he says

> consider the theory of light and color In our present terms, physics can be seen as an attempt to develop #theoretical entities# and #relations# whose structure is isomorphic to that of entities and relations in the real world If physics is right, for instance, the real world contains, among other things, electromagnetic radiation of various wavelengths , the properties of electromagnetic radiation are isomorphic to the #properties# of the theoretical construct #electromagnetic radiation# [29]

Cases in which both #-ed and #-less versions of expressions refer are extremely important in sustaining the two-worlds charge In such cases, reference is made to *both* (continuing with Jackendoff's example) electromagnetic radiation, in the theory of the hypothetical omniscient observer, and to electromagnetic-radiation-in-the-projection—that is #electro-magnetic radiation#—in the thinking of everyday thinkers [30] These cases emphasize that our inability to achieve reference to the real world is a result of pervasive cognitive

[28] *Semantics and Cognition*, pp 31–32 Even if, as Jackendoff suggests, the semantics of the theoretical language of psychology, and of science more generally—of Jackendoff's "metalanguage"—can be provided, it is problematic to hold that its semantics can be construed truth conditionally, given the spirit of Jackendoff's overall position Presumably the cognitive principles of our "science-forming faculty" (see numerous discussions in Chomsky, *New Horizons in the Study of Language and Mind* (New York Cambridge, 2000)) will shape the content of our scientific language and judgments in the same way that the principles of the perceptual system shape the content of perceptual experience But this initiates the move from truth to truth-in-the-projection for scientific language and judgments

[29] *Semantics and Cognition*, p 32

[30] These cases are also important in clarifying the semantics of '#' Cases in which #-less expressions do not refer, because there are no real world entities to refer to, are cases in which there are no real world correlates to projected entities (for example, the putative property of being red, on the assumption that there are no such things as colors, and thus no property of being red) These cases suggest that #-ed expressions are not semantically structured, for there may be no semantic values for #-less expressions from which to construct that semantics But cases where both a #-ed and its #-less counterpart expression refer show that that cannot be exactly right, for the unstructured semantics fails to capture the idea that the semantic value of a #-ed expression is the in-the-projection *correlate*, in some sense, of the semantic value of the #-less expression So it seems that the semantics for '#' has, somehow, to determine the value of the #-ed expression based on the value of the #-less expressions

limitation, and not a result of limitations imposed by a world that does not cooperate by providing referents

In any event, I will now use all the resources that Jackendoff has put on the table to see whether the problem of interpreting the instances of the T-schema can be overcome It might be objected on Jackendoff's behalf that he can interpret

(TI) 'snow is white' is true iff snow is white

in such a way that it is true, and obviously so, and further, that it is not (TI), but rather (TI#) that is the real target of the quick argument I gave three paragraphs back

(TI#) 'snow is white' is true iff #snow is white#

Here we utilize Jackendoff's distinction between #-ed and #-less expressions, and the fact that such expressions refer to different entities, indeed worlds (TI#), the objection continues, is no help with interpreting instances of the T-schema because (a) the expressions within the #-marks refer to snow- and being white-in the projection, and (b) mentioning snow- and being white-in–the-projection is not a way to state the *truth* condition for 'snow is white', but rather a way to state only the truth-in-the-projection condition More generally, using natural language sentences to provide the truth conditions for natural language sentences *must* fail, given that natural language sentences have at best truth-in-the-projection conditions (TI), the objection concludes, is subject to no such problem because the #-less expressions on the RHS do not refer to entities-in-the-projection and thus *can* serve to state genuine truth conditions

But this attempted recovery will not work The T-schema is not a claim about a form of words, orthographically typed

(O) $\ulcorner p \urcorner$ is true iff p,

where we have just to substitute for the $\ulcorner p \urcorner$s with the same words, orthographically-typed That would allow as an instance of the T-schema the sentence

(OI) 'snow is white' is true iff snow is white,

where the metalanguage is English*, in which the sentence type 'snow is white' used on the RHS means that snow is black, and it would disallow

(TIF) 'la neige est blanche' is true-in-French iff snow is white

where the RHS is used with its usual sense The problem with the objection being offered on Jackendoff's behalf is that it construes

instances of (T), like (T1), on the analogy of (O1), where the RHS of the biconditional is not the use of the sentence mentioned on the LHS (or of a translation of it) This *must* be so, because it is *precisely* Jackendoff's view that the sentences of natural language do not *have* truth conditions, and thus cannot be used *to specify* a truth condition [31] So if the RHS of a biconditional specifies a truth condition, it *cannot* be a use of the sentence mentioned on the LHS But that is just the general model instanced by (O1), where the mentioned sentence on the LHS and the used sentence on the RHS are orthographically type-identical but semantically type-distinct And, of course (O1) has absolutely *nothing* to do with the concept of truth

More explicitly, Jackendoff's view seems to be subject to the following dilemma either the RHS specifies a truth condition, in which case it does not involve a use of the appropriate sentence—of the sentence mentioned on the LHS, or the sentence used on the RHS *is* a use of the sentence mentioned on the LHS, in which case, that sentence cannot be used to state the condition for the *truth* of the sentence, but rather only a condition for the *truth-in-the-projection* of the sentence Either way, the theory fails to interpret instances of the T-schema as true, and obviously so, for the T-schema is grounded in the platitude that a sentence is *true* iff things are as a use of *it*, or a translation of *it*, says they are

I conclude that the semantically excessive interpretation is committed to the worst possible theory of truth

IV SEMANTIC EXCESS IN DEFLATIONISM

Deflationism generally construed involves two tenets first, that truth is to be understood through the T-schema, and second that the T-schema is all that is involved in understanding truth The first tenet gives the positive account of truth, the second captures the deflationary spirit of the view—the idea, in Horwich's words, that truth has a certain purity through its independence from other ideas, in particular from its independence from ideas in the theory of representation

Now, it is a common thought that a semantic theory involves the concept of truth—for example, in Davidson's proposal for the form of a semantic theory, where the theorems of the theory take the form

[31] It might be objected here that since the expression mentioned on the LHS is #-less, it has truth conditions, and thus can be used to state its own truth condition But, in the argument the mentioned sentence on the LHS is a sentence of a language that people actually speak, it is precisely these sentences to which Jackendoff's theory is to apply The argument can be extended to scientific sentences, as long as the role and bearing of principles of the "science-forming faculty" are kept in mind (see note 28)

of statements of truth conditions [32] These kinds of semantic theories are consistent with the thought that truth conditions *do not exhaust* meaning When truth conditions do not exhaust meaning, the relation between truth and meaning cannot be expressed with the slogan that a semantic theory *is* a theory of truth, but must rather must be expressed more carefully, as the thought that a semantic theory is *constrained* by a theory of truth, or more simply, as the thought that meanings are constrained by truth conditions This is the traditional use of truth in a semantic theory (for example, in the Fregean slogan that sense determines reference), and it is thus plausible to think that this is what the deflationist must deny in order to preserve the purity of truth More specifically, to accept the idea that truth is pure with respect to meaning seems to require denying both the stronger thesis that meanings *are* truth conditions and the weaker thesis that meanings are *constrained* by truth conditions It is only on an understanding of deflationism in which both the stronger and weaker theses about the connections between truth and meaning are denied that deflationism emerges as an interesting and controversial doctrine To allow that meanings are constrained by reference and truth conditions—that is, that meanings involve reference and truth conditions *as a rule*—is to link truth to meaning in an unacceptably strong way for one who wishes truth to be pure It is hard to see how that kind of tight linking can have no inflationary effect on the concept of truth [33]

I think that this understanding of deflationism is common ground between me and at least one very influential and ambitious deflationist, namely Hartry Field [34] Field discusses what he calls "two traditions" of thinking about meaning those in which meaning and content "encapsulate" truth conditions, and those that do not Field

[32] See Davidson (*op cit*)

[33] Depending upon how inflated or deflated the account of meaning is For a discussion of these themes that sets them in the contexts of some recent philosophical history and some basic questions in the foundations of semantics, see my "The Theory of Truth in the Theory of Meaning," *European Journal of Philosophy*, XII (2004) 219–43, and "Davidson, Semantic Deflationism, and Dummett's Dilemma," *Iyyun*, LIV (2005) 39–63, respectively Some deflationists, like Horwich (*op cit*) and Williams (*op cit*), allow that deflationism about truth is consistent with accounts in which meaning is constrained by reference I am for the most part not concerned with these weaker, less pure, versions of deflationism, although I briefly consider these views in replying to objection (iv) in section v below (So, despite having initiated my discussion with Horwich's pithy remark, I do not think that his view conforms to a standard of purity that makes deflationism a deeply controversial thesis—cf Field, "Critical Notice Paul Horwich's *Truth*," pp 326–27)

[34] Field, "Deflationist Views of Meaning and Content" (1994), reprinted in Field, *Truth and the Absence of Fact* (New York Oxford, 2001), pp 104–40

urges the deflationist about truth to be sure that the account of *meaning* that she will eventually need to provide does not "encapsulate" a *reduction* of truth conditions and reference

> If deflationism is to be at all interesting, it must claim not merely that what plays a central role in meaning and content not include truth conditions *under that description*, but that it not include *anything that could plausibly constitute a reduction of* truth conditions to more physicalist terms (*ibid*, p 108)

Field's entirely correct idea is that a deflationary account of *meaning* that serves to reduce or even partially reduce truth conditions would serve to produce a very inflationist view of *truth* Meanings that determined reference and truth conditions would be at exactly such a risk of inflating truth So, for deflationism about meaning to be an interesting doctrine, it must not allow that meanings are constrained by truth conditions Holding truth and meaning apart like that will guard the purity of truth and ensure that that one's account of meaning does not inflate one's account of truth

Field takes these points very seriously Field's view of meaning is that meaning facts are grounded in facts about the *uses* of expressions—in particular in facts about the conceptual role and indication relations of expressions Following Field, I will call these properties of the uses of expressions their *meaning characteristics* Field goes out of his way to show that meanings construed as such could not be taken to reduce, even partially, the notion of reference or truth conditions, for in many cases the meaning characteristics of sentences *diverge from*, and do not merely *go beyond*, the truth conditions of those sentences Field approaches the point by saying that the role of meaning characteristics, and in particular indication relations

> might make one think that the deflationist is bound to recognize (a non-disquotational version of) the relation "S has truth conditions *p*" in fact if not in name, since these indication relations constitute the truth conditions relation But this overlooks the fact that the project of giving anything close to a believable reduction of talk of truth conditions to talk of indication relations is at best a gleam in the eye of some theorists A way to see the point is to notice that there are plenty of examples where the indication relations don't reflect what we would intuitively regard as truth conditions (*ibid*, p 110)

Talk of reduction may indeed be implausible, but Field's real point here is to provide a response to the worry about whether meaning characteristics inflate truth Meaning characteristics cannot inflate truth because the account of meaning in terms of meaning characteristics actually *conflicts* with accounts in terms of reference and truth

conditions [35] To use an example of Field's, although the truth conditions of 'In Bosnia, *p*' have something to do with *p*-type states of affairs in Bosnia, the meaning characteristics and in particular the indication relations of the sentence may reflect only what is said in certain newspapers, where what is said in these newspapers indicates either nothing systematic or something other than what is happening in Bosnia So the meaning characteristics make the sentence semantically connected to what is said in the newspapers, rather than about what is taking place in Bosnia.

In general, these extensional differences between meaning characteristics and reference and truth conditions shows that meaning characteristics cannot be a reductive specification of reference and truth conditions

So far, I have been arguing that the only interesting form of *truth-theoretic* deflationism is one according to which the account of *meanings* is not constrained to determine truth conditions But, to continue with the Bosnia example, that is emphatically *not* to say that there are no *p*-type states of affairs obtaining, or failing to obtain, in Bosnia It is not to say that there is no Bosnia, and only newspapers It is precisely these *p*-type states of affairs and Bosnia that are mentioned in a statement of the truth conditions for 'In Bosnia, *p*' The point is just that the meaning characteristics *do not* determine those *p*-type states of affairs obtaining in Bosnia, and do not determine Bosnia, as having any relevance, semantically, for 'In Bosnia, *p*' Only those things that are determined by the meaning characteristics of the sentence are semantically relevant to the sentence [36]

Let me regiment Field's distinction between on the first hand, what is determined by the meaning characteristics of a sentence, and on the other, "what we would intuitively regard as truth conditions" for the sentence Let us stipulate that *propositions* are those things that are

[35] That is, Field is not merely despairing over the possibility of providing a naturalistic reduction of reference and truth conditions, although that kind of despair probably does inaugurate his move from causal theorist of reference in "Tarski's Theory of Truth" (reprinted in Field, *op cit*, pp 3–26) to deflationist in "Deflationist Views of Meaning and Content" (cf the Preface of *Truth and the Absence of Fact*) The point is rather that if such a reduction were in the cards with the account of meaning characteristics, then the position would cease to be deflationary

[36] And these very well may be things, properties, and states of affairs in the world So, the objection here will not be that the deflationary view of truth "cuts language off from the world" ("Deflationist Views of Meaning and Content," p 126) The objection recognizes that deflationism does make use of meaning characteristics, especially indication relations, to recover the connection between language and the world That is *not* the objection that I will be pursuing The objection that I will be pursuing revolves around the separation of representation from reference and truth conditions

intuitively expressed by sentences, things that are individuated by "what we would intuitively regard as truth conditions " So for example, 'snow is white' intuitively expresses the proposition *that snow is white,* and both the sentence and the proposition are true iff snow is white In this terminology, we can now rephrase my, and Field's, point about the only controversial and interesting form that the deflationary thesis can take, as follows that the meaning characteristics of expressions, and sentences in particular, should not be constrained to determine propositions, in this stipulated sense Whatever sentences intuitively express, this is not what they, as a semantic matter, express So, in whatever sense of 'meaning' meaning characteristics metaphysically determine meanings—let us say they determine *deflated meanings* or *deflated propositions*—they cannot determine propositions, that is, they cannot determine those things that are intuitively expressed by sentences and are individuated by "what we would intuitively regard as truth conditions "

So, on one hand we have propositions, and on the other, deflated propositions The deflationist holds that our sentences express deflated propositions, and not propositions This separation allows the deflationist theory of meaning in terms of meaning characteristics to be such as not to inflate truth [37]

For reasons that should be starting to become apparent, I want to regiment this view further by using expressions surrounded by #-marks to *intuitively express* whatever is metaphysically determined by meaning characteristics So, for example, '#snow is white#' serves where 'snow is white' could not 'Snow is white' does not intuitively express what is metaphysically determined by meaning characteristics, but intuitively expresses a proposition individuated by "what we would intuitively regard as truth conditions " But '#snow is white#' *does* intuitively express whatever is determined by meaning characteristics—that is, it expresses a deflated proposition, one that is distinct from propositions individuated by "what we would intuitively regard as truth conditions "

The set-up has been (really!) long and laborious, but the argument against deflationism should now be clear it is subject to an argument exactly parallel to that offered against the semantically excessive

[37] I have some reservations about attributing this position to Field himself, although, as Field also reports ("Deflationist Views of Meaning and Content," p 111, note 9), it is a position that one often encounters from proponents of deflationism Although I have recounted Field's view in some detail here, I am unsure of what it looks like in its finest detail See the responses to objections (iii) and (iv) in section v below for further relevant discussion This is also a good place to reiterate that the focus of my critique is the philosophy behind deflationism, and not any deflationist in particular

interpretation What will the deflationist say about the instances of the T-schema? Once again, there seems to be no interpretation of a T-sentence in which it comes out as true and obviously so, neither

(TI) 'snow is white' is true iff snow is white

nor

(TI#) 'snow is white' is true iff #snow is white#

will do Remember, that for the deflationist, the sentences mentioned on the LHS do not express propositions, but instead express deflated propositions, and these are by design distinct from propositions Consider now (TI) In (TI) the RHS expresses "what we would intuitively regard as truth conditions" for the sentence 'snow is white', but by dint of precisely this fact, it cannot be a use of the sentence mentioned on the LHS So (TI) does not conform to (Plat), according to which a sentence is true iff things are as a use of *it*, or a translation of *it*, says they are What about (TI#)? By the deflationist lights, the RHS is a use of the sentence mentioned on the LHS (or better, a use of a translation of it), but, by now by dint of that very fact, it cannot be the use of a sentence that expresses "what we would intuitively regard as truth conditions" for the sentence mentioned on the LHS So neither (TI) nor (TI#) interpret the instances of the T-schema in such a way that come out true, and obviously so

I am about to go on and consider objections and offer replies, but one objection needs to be considered and responded to right away It does the deflationist no good to say that unlike Jackendoff and the proponent of the semantically excessive interpretation, her sentences do speak of the world, but just not the bits of the world that, intuitively, the sentences to which truth is ascribed seem to speak about Put another way, the deflationist may say that unlike the proponent of the semantically excessive view, her sentences do express truth conditions, just not "what we would intuitively regard as truth conditions " But what does that mean? The deflationist *cannot* mean by this that her sentences express the *real* truth conditions for sentences, and that "what we would intuitively regard as truth conditions" are not the real truth conditions For example, the deflationist cannot mean that the real truth conditions of 'In Bosnia, *p*' make mention of newspapers and the like, and not of *p*-type states of affairs in Bosnia For that would just be a view in which meaning *did* determine truth conditions, and a view in which the meaning characteristics of sentences *did not* come apart from the truth conditions of sentences, and a view in which meaning characteristics may very well comprise the air that inflates the account of truth If that were the view, then deflationism

would not be a *controversial view of our understanding of truth and meaning*, but a controversial view about what our words mean and the conditions under which sentences containing them are true, understood against the background of an *uncontroversial view of our understanding of truth and meaning* What the deflationist *must* mean by saying that her sentence expresses a truth condition but not "what we would intuitively regard as truth conditions" for the sentence mentioned on the LHS is that it expresses *a* truth condition, but not *the* truth condition for the sentence, because whatever it is that '#snow is white#' expresses, it cannot be *the* truth condition for 'snow is white' because it was engineered by the deflationist to express, precisely, *anything but that*[38]

Let me, in closing this section, state the objection to deflationism in a way that more clearly parallels that offered against the semantically excessive interpretation, but that also incorporates the points I have just made If the sentence used on the RHS specifies "what we would intuitively regard as truth conditions" for the sentence mentioned on the LHS, it cannot be a use of the sentence mentioned on the LHS, for, according to the deflationist, that sentence does not have a meaning that determines "what we would intuitively regard as truth conditions" for the mentioned sentence On the other hand, if the sentence used on the RHS *is* a use of the sentence mentioned on the LHS, it cannot be used to state *the* truth condition for the mentioned sentence even if it can be used to state *a* truth condition, for the meaning characteristics of the mentioned sentence, used on the RHS, do not determine *the* truth condition Either way, the theory fails to interpret instances of the T-schema in such a way that they are true, and obviously so

The two-worlds structure of a view consists of the separation of representation from truth in that view A two-worlds structure is present in a relatively obvious way in the semantically excessive interpretation (section II) The presence of the two-worlds structure in a view leads proponents of that view to be committed to the worst possible theory of truth (section III) The two-worlds structure is also present in deflationism (section IV) I conclude that the deflationist, too, is committed to the worst possible theory of truth

[38] The two-worlds structure of the semantically excessive interpretation does not come from its commitments to a mental, projected world and the real world, but rather from its separation of representation from truth and attachment to some notion that is explicitly distinguished from truth The problem is the distinction between a world of something-other-than-truth in representation and the world of truth that is disconnected from representation

V SOME OBJECTIONS AND REPLIES

At this point, one might think that either the *characterization* of deflationism is unfair, or that the *connection* between the deflationary accounts of truth and meaning has been misconstrued In particular, it may be objected that the characterization and connection are not consistent with the following aspects of deflationism (i) that the T-schema is stipulated to be true by convention and to fix the reference of 'true', (ii) that the cognitive rule that is to be followed to possess the concept of truth is the simple and underived one of having a disposition to judge the instances of the T-schema, (iii) that meaning characteristics are not part of an account of what meaning consists in, and (iv) that deflationism need not involve the radical disconnect between truth and meaning that I have arrived at in the discussion of Field I will respond to these objections one at a time

(i) First, consider the idea that the instances of the T-schema are stipulated to be true by convention It may well be the case that the truth of an instance of the T-schema, at least for linguistic versions of the T-schema, is conventional But this will be conventional truth only in the sense that the truth is a truth *about* certain (linguistic) conventions, not a truth *by* convention But, as propositionalist versions of the T-schema show, the instances of the T-schema need not be conventional even in this weak sense of being truths about (linguistic) conventions An alternative view is that the instances of the T-schema comprise conceptual knowledge about the property of being true But this need not commit one to any kind of conventionalism Consider in this regard W V O Quine's criticisms of a version of the analytic/synthetic distinction in which the analytic truths are made true by conventionally determined meanings and the synthetic truths by the world [39] What the criticisms show is that the distinction between analytic and synthetic is not a distinction in domains of truth makers What makes analytic or conceptual truths *conceptual* is that standing in some relation to them, like judging or knowing them, is fundamental to answering the question of what it is to be able to think in a certain way about some aspect of the world What makes them true is the world Applied to the T-schema, the view is that the instances of the T-schema are such that standing in some relation to them, like judging or knowing them, is fundamental to the issue of what it is to be able to think in a certain way about the property of being true On this view, the instances of the T-schema are made conceptual not by a special domain of conventionally determined truth

[39] Quine, "Carnap and Logical Truth" (1956), reprinted in his *Ways of Paradox and Other Essays* (Cambridge Harvard, 1980, 2nd ed), pp 100–25, §3

makers but by the role of the T-schema in answering the question of what it is to be able to think in a certain way about aspects of the world, in this case the property of being true What makes the instances of the T-schema true are facts about the property of being true

These considerations together with the myriad other problems that attend truth by convention suggest that deflationism would do well to distance itself from the very problematic idea of truth by convention

(ii) A second, more important, objection criticizes the idea that the deflationist theory of meaning or content for 'snow' or *snow* has any impact or consequences for the deflationist account of truth It cannot have any impact or consequences because there is no room for it to do so, and there is no room for it to do so because the possession conditions for the concept of truth are simple, and consist (roughly) in having a disposition to accept or judge the instances of the T-schema But, the objection continues, the view here distorts this idea by making possession of the concept of truth depend upon other concepts, and this makes the possession condition no longer simple

The objection confuses the potentially correct idea that the possession conditions of the concept of truth are simple, with the incorrect idea that simple possession conditions can make no *further* demands on the conceptual repertoire of the thinker But a more general consideration of implicit definitions, from which this account of possession of the concept of truth draws inspiration (where the instances of the T-schema implicitly define truth) shows that this idea is mistaken An implicit definition gives the meaning of some expression by specifying its contribution to the truth conditions of canonical sentential contexts (or a canonical sentential context schema) in which the expression figures But a canonical *sentential* context will mention other expressions, and accepting the truth of a canonical sentential context and determining the correct reference for the expression being implicitly defined will depend on the concepts expressed by and references of these other expressions So although it may be the case that possessing the concept of truth requires accepting instances of the T-schema, all the expressions in an instance and their semantics are relevant to whether the reference being determined for the expression being defined—'true'—is the correct one The objection of this paper is that the deflationist account of meaning makes the canonical sentential contexts for 'true'—the instances of the T-schema—useless for implicitly defining and determining the reference of 'true' because the instances will not conform to (Plat) As an analogy, consider a view according to which the truth values of sentences that express numerical identities between *F*s and *G*s (the number of *F*s = the number of *G*s) are determined by the condition that the *F*s and *G*s can be put into

one-to-one correspondence, but which is then followed by an account of one-to-one correspondence in which it is not an equivalence relation No one would be inclined to say that this version of Hume's Principle plays a role in our understanding of number

(iii) The third objection criticizes the idea, implicit in my presentation, that the account in terms of meaning characteristics is an account of what meaning consists in for the deflationist That is, it might be objected that meaning characteristics are not to be understood as contributing to an account of what meaning consists in [40] Perhaps Field's point in emphasizing how it is that the meaning characteristics of a sentence can conflict with its intuitive truth conditions is that the meaning characteristics *cannot* be an account of what meaning facts consist in Moreover, Field holds that whether a sentence is true or not is a use-*independent* fact about the sentence, and it would be odd, then, to hold that the meaning of sentences is a use-*dependent* property of the sentence The only appropriate theory of meaning would mimic the deflationary theory of truth in structure, and hold the meanings of sentences to be determined by a schema like

(M) $\ulcorner S \urcorner$ means that S,

and not by anything more substantive But, the objection continues, since the arguments offered here work by highlighting a tension between the deflationist theory of truth and the deflationist theory of meaning, these objections will get no traction if the account in terms of meaning characteristics is not an account of meaning for the deflationist

I am happy to admit that the objections of this paper do not apply if the account in terms of meaning characteristics is not an account of meaning for the deflationist [41] I can only offer the following comments about the proposal First, it is not clear at all what meaning characteristics are if not the characteristics that determine the facts about meaning Second, even if an account in terms of meaning characteristics is not supposed to be an account of meaning for the deflationist, there remains the question of what the deflationist account of meaning is to be Finally, if the account of meaning is not provided by meaning characteristics, but is instead provided by the extremely deflated schema (M) and its instances, it will be subject to numerous problems of its own, including one that requires an answer

[40] This is not an objection somebody like Horwich could avail himself It is very clear that his use theory of meaning is an account of what constitutes meaning

[41] See my "Davidson, Semantic Deflationism, and Dummett's Dilemma," §5

to the motivational question of why we should rest content with such a modest account of meaning [42]

(iv) The fourth objection says that although the argument of this paper may be cogent against the very radical kind of deflationism that I have arrived at in the discussion of Field, it does nothing to harm less radical versions Does truth really need to be *that* pure with respect to meaning in order to be deflated? The idea here is that truth can play a role in the theory of meaning—in particular, the role of constraining meanings, without compromising a deflationary account of truth [43] Suppose that truth is understood through the instances of the T-schema, and that truth plays a role in constraining meaning, and in this sense, meaning is truth conditional Would that not be deflated enough?

This is not the place to pursue the issue in any significant detail Others have argued that deflationism is not consistent with a truth conditional account of meaning, because the combination involves a vicious circularity [44] My own view is that the problem is not exactly circularity, but rather that the proximity into which a truth conditional account of meaning brings the concepts of meaning and truth is such as to inflate truth to the same extent that the account of meaning is inflated (see note 33) But for current purposes, I am inclined just to turn this objection on its head My main concern is to argue against deflationism as a philosophy of how to understand truth, meaning, and their interrelations The kind of deflationism that I have arrived at in the discussion of Field is a radical philosophy of truth, meaning, and their interrelations Deflationist views that are consistent with truth-conditional accounts of meaning do not have the kinds of radical repercussions for our understanding of truth, meaning, and their interrelations that is the focus of my interest (this is not to say that they have no repercussions) Paraphrasing Davidson

[42] Here the considerations do not parallel matters in the theory of truth We may, perhaps, answer the motivational question for deflationism about truth by showing how deflated truth can do the work we want—say, logical work—of substantial truth But we do not want an account of meaning because of the work it can do for us, we want it because we want to know *what meaning is* Motivating deflationism about meaning is hard McDowell, for example, favors a deflated account of meaning—what he calls a *modest* account of meaning—and goes to great lengths to justify the idea that an account of meaning should not be anything more than modest, see "In Defence of Modesty", and also "Another Plea for Modesty" (1997), reprinted in his *Meaning, Knowledge, and Reality*, pp 108–31 For criticism, see my "The Theory of Truth in the Theory of Meaning"

[43] Horwich, *op cit*, Williams, *op cit*

[44] Most notably Michael Dummett, "Truth" (1959), reprinted in his *Truth and Other Enigmas* (Cambridge Harvard, 1978), pp 1–24 For some criticisms of this idea, see Patterson, "On the Determination Argument against Deflationism," *Pacific Philosophical Quarterly*, LXXXVIII (2007) 243–50, and Claire Horisk, "Truth, Meaning, and Circularity," *Philosophical Studies* (forthcoming)

from a not unrelated context, such views may improve intelligibility, but not while retaining the excitement [45]

VI ON THE VERY IDEA OF THE T-SCHEMA

My strategy of argument against deflationism has been to associate deflationism with the semantically excessive interpretation, and in particular to reveal its common commitment to a semantically excessive understanding of representation This leads both the semantically excessive interpretation and deflationism to the worst possible theory of truth However, I think that there are at least two interrelated points of disanalogy between the semantically excessive interpretation and deflationism Although I doubt it, these disanalogies may hold out further hope for the semantically excessive interpretation, but they further cement the case against the deflationist

The first disanalogy is the following My argument charges the semantically excessive interpretation of failing to provide an account of meaning that is in conformity with a fundamental platitude that links the concepts of meaning and truth It might be replied that my argument cannot harm the kind of scientific project that Jackendoff is engaged in, or the conclusions that he arrives at, because scientific projects do not deploy *common sense* concepts, but rather *technical* concepts, and these technical concepts are not constrained by the intuitive platitudes conformity with which partially characterizes the commonsense concepts [46] I think the objection is mistaken, because the argument does not intend to put the scientific project, but only the interpretive gloss that that scientific project is given, into jeopardy—this is why the problem concerns not cognitive science, but a semantically excessive interpretation of it Be that as it may, I want to note now that this kind of response is simply not open to the

[45] Davidson, "On the Very Idea of a Conceptual Scheme" (1974), reprinted in his *Inquiries into Truth and Interpretation*, pp 183–98, at p 183 The ideas of this paragraph can also be used to understand a progression in Davidson's thought, from the view that a deflated concept of truth functions as a theoretical primitive in the theory of meaning to the view that the concept of truth is substantially explicated by the theory of meaning This movement away from deflationism in Davidson is not based in the objection, criticized by Kirk Ludwig ("Davidson's Objection to Horwich's Minimalism about Truth," this JOURNAL, CI, 8 (August 2004) 429–37), that no compositional sense can be made, within the confines of deflationism, of the semantics of 'the proposition that *p*' (see Davidson, "The Folly of Trying to Define Truth," this JOURNAL, XCIII, 6 (June 1996) 263–78) Rather, the movement away from deflationism conforms to the general idea, in Ludwig's words, that "[o]bjections to minimalism must rest on establishing important connection between the concept of truth and other concepts" (p 437)

[46] An argument against referential or truth conditional semantics explicitly along these lines can be recovered from Chomsky, *New Horizons in the Study of Language and Mind*, chapter 2

deflationist The deflationist is staking out a position on the concepts of truth and meaning that is meant to quiet the debates about concepts that are, and have forever been, a central concern of philosophy Those debates surely do not concern technical concepts of recent vintage It concerns our folk understanding of a concept that is central to our dealings with the world and with others [47]

The second, related, disanalogy is that it might be replied on Jackendoff's behalf, but not the deflationist's, that it is *precisely* his view that sentences *do not* have truth conditions—they have only truth-in-the-projection conditions—and thus, to say that the position is unable to recover the connection between meaning and truth that is embodied in (Plat) is simply to make explicit a central feature of the intended, technical, replacement that is being offered The criticisms offered here, continues the objection, simply beg the question I think that this objection mistakes the character of the argument given here The argument does not simply confront a non-truth conditional account of meaning with a truth conditional one, as I said earlier, my objections do not apply to other criticisms of truth conditional semantics Rather, the structure of the argument is to note that Jackendoff's form of commitment to a non-truth conditional theory of meaning entails absurdity in one's theory of truth, it entails that the instances of the T-schema cannot be interpreted in any way in which they come out true, and obviously so In any event, whatever the merits of this objection, it should be noted that it is available to the deflationist only on pain of incoherence We have just pointed out that deflationism is not concerned with technical replacements for certain central folk concepts, but with those central folk concepts themselves But the account that deflationism provides for our folk concept of truth is precisely that which is embodied in the list of instances of the T-schema, and this list serves to characterize truth only in virtue of the conformity of the instances to (Plat) But the fact that deflationary meaning characteristics do not determine truth conditions means that the deflationary view of truth—the entire positive dimension of the deflationist position—itself has been undercut, it cannot understand the instances of the T-schema as instances of a schema that conforms to the platitude about truth and meaning that a mentioned sentence is true iff things are as a use of it say they are

[47] For an illuminating discussion of whether Field's deflationary accounts of meaning and truth can serve as either an analysis of our ordinary notions, or as a technical replacement, see Ray Buchanan, "Are Truth and Reference Quasi-Disquotational?" *Philosophical Studies*, CXIII (2003) 43–75

CENTRAL LIBRARY UNIVERSITY OF CALCUTTA

To put the point explicitly: the argument against Jackendoff has the structure of a *reductio*, but in addition to being subject to an exactly parallel argument, the deflationist is also subject to a charge of internal inconsistency or incoherence [48]

GURPREET RATTAN

University of Toronto

[48] I think that this point is missed in Dorit Bar-On, Horisk, and William Lycan, "Deflationism, Meaning, and Truth Conditions," *Philosophical Studies*, CI (1999) 1–28, in their interesting account of the problems that deflationism faces when it comes to providing an account of meaning They aim to establish in their paper that the demands placed on truth by meaning entail that "either deflationism about truth is false or the received view that deflationism about truth is incompatible with a truth conditional account of view of meaning must be rejected" (p 1) But, on my view, the demands placed on truth by meaning render deflationism incoherent. if meaning is not truth conditional, then the deflationist both accepts and denies the instances of the T-schema, and if meaning is truth conditional, then the deflationist both accepts and denies that there is more to say about truth than that which is captured in the instances of the T-schema

THREE KINDS OF IDEALIZATION*

Philosophers of science increasingly recognize the importance of *idealization* the intentional introduction of distortion into scientific theories Yet this recognition has not yielded consensus about the nature of idealization The literature of the past thirty years contains disparate characterizations and justifications, but little evidence of convergence towards a common position

Despite this lack of convergence, consensus has clustered around three types of positions, or three kinds of idealization While their proponents typically see these positions as competitors, I will argue that they actually represent three important strands in scientific practice Philosophers disagree about the nature of idealization because there are three major reasons scientists intentionally distort their models and theories, all three kinds of idealization play important roles in scientific research traditions

The existence of three kinds of idealization means that some classic, epistemic questions about idealization will not have unitary answers We cannot expect a single answer to questions such as What exactly constitutes idealization? Is idealization compatible with realism? Are idealization and abstraction distinct? Should theorists work to eliminate idealizations as science progresses? Are there rules governing the rational use of idealization, or should a theorist's intuition alone guide the process? However, the three kinds of idealization share enough in common to allow us to approach the answers to these questions in a unified way The key is to focus not just on the practice and products of idealization, but on the goals governing and guiding it I call these goals the *representational ideals* of theorizing Although they vary between the three kinds of idealization, attending to them will help us better understand the epistemic role of this practice

* Many thanks to Peter Godfrey-Smith, Stephan Hartmann, Paul Humphreys, Steve Kimbrough, Ryan Muldoon, Michael Strevens, Ken Waters, and Deena Skolnick Weisberg for extremely helpful comments and advice I am also grateful for the thoughtful questions and comments from audiences at the Minnesota Center for Philosophy of Science, Tilburg University, the University of Pennsylvania, and Washington and Lee University, where earlier versions of this paper were presented Thanks also to students in my graduate seminars at Penn for ongoing, stimulating discussion about idealization The research in this paper was partially supported by National Science Foundation grant SES-0620887

0022-362X/07/0412/639–59 © 2007 The Journal of Philosophy, Inc

I THREE KINDS OF IDEALIZATION

Since the early 1980s, philosophers of science have paid increasing attention to the importance of idealization in scientific inquiry While earlier literature acknowledged its existence, the pioneering studies of Nancy Cartwright,[1] Ernan McMullin,[2] Leszek Nowak,[3] William Wimsatt,[4] and others paved the way for the contemporary philosophical literature on the topic Through much of my discussion, I will follow Cartwright's characterization and talk about theoretical representation in terms of *modeling*, the indirect representation of real world phenomena with models [5] But many of the ideas in this paper are not essentially tied to modeling, so my reliance on the model-based idiom should not be seen as affirming this connection

One of the most important insights of the modern idealization literature is that idealization should be seen as an activity that involves distorting theories or models, not simply a property of the theory-world relationship This suggests that in order to distinguish between the three types of idealization we will need to know what activity is characteristic of that form of idealization and how that activity is justified These activities and justifications can be grouped into three kinds of idealization *Galilean idealization*, *minimalist idealization*, and *multiple-models idealization*

Galilean Idealization Galilean idealization is the practice of introducing distortions into theories with the goal of simplifying theories in order to make them computationally tractable One starts with some idea of what a nonidealized theory would look like Then one mentally and mathematically creates a simplified model of the target

Galilean idealization has been thoroughly characterized and defended by McMullin who sees the point of this kind of idealization as "grasp[ing] the real world from which the idealization takes its origin," by making the problem simpler, and hence more tractable (*op cit*, p 248) [6] Galileo employed the technique both in theoretical

[1] Cartwright, *How the Laws of Physics Lie* (New York Oxford, 1983), and *Nature's Capacities and Their Measurements* (New York Oxford, 1983)

[2] McMullin, "Galilean Idealization," *Studies in History and Philosophy of Science*, XVI (1985) 247–73

[3] Nowak, "Laws of Science, Theories, Measurement," *Philosophy of Science*, XXXIX (1972) 533–48

[4] Many of Wimsatt's most important papers on idealization and related topics are collected in Wimsatt, *Re-engineering Philosophy for Limited Beings Piecewise Approximations of Reality* (Cambridge Harvard, 2007)

[5] For more detail about the practice of modeling, see Michael Weisberg, "Who Is a Modeler?," *British Journal for the Philosophy of Science*, LVIII (2007) 207–33

[6] A similar account is developed by Nowak, see Nowak, "The Idealizational Approach to Science A Survey," in Jerzy Brzezinski and Nowak, eds , *Idealization III Approximation*

and experimental investigations Although this paper is concerned with the former, Galileo's vivid description of the experimental version is useful for conceptualizing the basic notion of Galilean idealization When discussing the determination of gravitational acceleration in the absence of a medium devoid of resistance, Galileo suggests a kind of experimental idealization

> We are trying to investigate what would happen to moveables very diverse in weight, in a medium quite devoid of resistance, so that the whole difference of speed existing between these moveables would have to be referred to inequality of weight alone Since we lack such a space, let us (instead) observe what happens in the thinnest and least resistant media, comparing this with what happens in others less thin and more resistant [7]

Lacking a medium devoid of resistance, Galileo suggests that we can make some progress on the problem by initially using an experimental setup similar to the envisioned situation After understanding this system, the scientist systematically removes the effect of the introduced distortion The same type of procedure can be carried out in theorizing introduction of distortion to make a problem more tractable, then systematic removal of the distorting factors

Galilean idealization is justified pragmatically We simplify to more computationally tractable theories in order to get traction on the problem If the theorist had not idealized, she would have been in a worse situation, stuck with an intractable theory Since the justification is pragmatic and tied to tractability, advances in computational power and mathematical techniques should lead the Galilean idealizer to de-idealize, removing distortion and adding back detail to her theories With such advances, McMullin argues, "models can be made more specific by eliminating simplifying assumptions and 'de-idealization', as it were The model then serves as the basis for a continuing research program" (*op cit*, p 261) Thus the justification and rationale of Galilean idealization is not only pragmatic, it is highly sensitive to the current state of a particular science

Galilean idealization is important in research traditions dealing with computationally complex systems Computational chemists, for example, calculate molecular properties by computing approximate wavefunctions for molecules of interest At first, all but the simplest

---

and Truth, Volume 25 of *Poznan Studies in the Philosophy of the Sciences and the Humanities* (Atlanta, GA. Rodopi, 1992), pp 9–63

[7] Quoted in McMullin, p 267

systems were intractable When electronic computers were introduced to computational chemistry, calculated wavefunctions remained crude approximations, but more complex, chemically interesting systems could be handled As computational power has continued to increase in the twenty-first century, it has become possible to compute extremely accurate (but still approximate) wavefunctions for moderate sized molecules Theorists in this tradition aim to develop ever better approximations for molecular systems of even greater complexity [8] These techniques are still approximate, but research continues to bring computational chemists closer to the goal of "[calculating] the exact solution to the Schrödinger equation, the limit toward which all approximate methods strive "[9]

This example nicely summarizes the key features of Galilean idealization The practice is largely pragmatic, theorists idealize for reasons of computational tractability The practice is also nonpermanent Galilean idealization takes place with the expectation of future deidealization and more accurate representation

Minimalist idealization Minimalist idealization is the practice of constructing and studying theoretical models that include only the core causal factors which give rise to a phenomenon Such a representation is often called a *minimal model* of the phenomenon Put more explicitly, a minimalist model contains only those factors that *make a difference* to the occurrence and essential character of the phenomenon in question

A classic example of a minimalist model in the physical sciences is the Ising model This simple model represents atoms, molecules, or other particles as points along a line and allows these points to be in one of two states Originally, Ernst Ising developed this model to investigate the ferromagnetic properties of metals It was further developed and extended to study many other phenomena of interest involving phase changes and critical phenomena The model is powerful and allows qualitative and some quantitative parameters of substances to be determined But it is extremely simple, building in almost no realistic detail about the substances being modeled What it seems to capture are the interactions and structures that

[8] There are principled reasons why the exact wavefunction for multi-electron systems can not be computed However, there are no general, in-principle reasons why approximations of arbitrarily high degrees of accuracy and precision cannot be computed

[9] J B Foresman and A Frisch, *Exploring Chemistry with Electronic Structure Methods* (Pittsburgh Gaussian, 1996), p 95 For a discussion of the relevant philosophical issues, see Paul Humphreys, "Computer Simulation," in Arthur Fine, Micky Forbes, and Linda Wessels, eds , *Philosophy of Science Association 1990, Volume 2* (East Lansing, MI PSA, 1992), pp 597–609, and *Extending Ourselves* (New York Oxford, 2004)

really make a difference, or the core causal factors giving rise to the target phenomenon

Among recent discussions of idealization in the philosophical literature, minimalist idealization has been the most comprehensively explored position As such, there is some diversity among the articulations of this position One view is Michael Strevens's *kairetic* account of scientific explanation Strevens's account of explanation is causal, to explain a phenomenon is to give a causal story about why that phenomenon occurred What makes Strevens's account distinct is that the explanatory causal story is limited to only those factors that made a difference to the occurrence of the phenomenon "Making a difference" is a fairly intuitive notion, but Strevens defines it explicitly in terms of what he calls "causal entailment,"[10] which involves logical entailment in a causal model A causal factor makes a difference to a phenomenon just in case its removal from a causal model prevents the model from entailing the phenomenon's occurrence A causal model of the difference-making factors alone is called a *canonical explanation* of the target phenomenon

For Strevens, idealization is the introduction of false but nondifference-making causal factors to a canonical explanation In explaining Boyle's law, for example, theorists often introduce the assumption that gas molecules do not collide with each other This assumption is false, collisions do occur in low-pressure gases However, low-pressure gases behave as if there were no collisions This means that collisions make no difference to the phenomenon and are not included in the canonical explanation Theorists' explicit introduction of the no-collision assumption is a way of asserting that collisions are actually irrelevant and make no difference [11] Even with this added, irrelevant factor, the model is still minimalist because it accurately captures the core causal factors

Other accounts of minimalist idealization associate minimalism with generation of the canonical explanation alone Robert Batterman's account of asymptotic explanation is an example of such a view Asymptotic methods are used by physicists to study the behavior of model systems at the limits of certain physical magnitudes These methods allow theorists to study how systems would behave when certain effects are removed, which allows the construction of "*highly idealized minimal models* of the universal, repeatable features of a sys-

[10] Strevens, "The Causal and Unification Accounts of Explanation Unified—Causally," *Noûs*, XXXVIII (2004) 154–76

[11] Strevens, "Why Explanations Lie Idealization in Explanation" (unpublished manuscript, September 2004), p 26

tem "[12] These minimal models have a special role in physics because they can be used to explain universal patterns, common behaviors across material domains such as pressure, temperature, and critical phenomena Adding more detail to the minimal model does not improve the explanations of these patterns, more details only allow a more thorough characterization of a highly specific event

Arguing in a similar vein, Stephan Hartmann describes cases where highly complicated systems are characterized using physical models "of (simple) effective degrees of freedom," which help to give us "partial understanding of the relevant mechanisms for the process under study" This plays a cognitive role by allowing theorists "to get some insight into the highly complicated dynamics" of such systems [13]

Cartwright's account of *abstraction* is also an example of what I call minimalist idealization [14] On her view, abstraction is a mental operation, where we "strip away—in our imagination—all that is irrelevant to the concerns of the moment to focus on some single property or set of properties, 'as if they were separate' " If the theorist makes a mathematical model of this abstract, real phenomenon, then she is in possession of a minimal model Such a model can reveal the most important causal powers at the heart of a phenomenon [15]

Despite the differences between minimalist idealization and Galilean idealization, minimalist idealizers could in principle produce an identical model to Galilean idealizers For example, imagine that we wanted to model the vibrational properties of a covalent bond A standard way to do this is to use a harmonic oscillator model This model treats the vibrating bond as spring-like with a natural vibrational frequency due to a restoring force This is a very simple representation

[12] Batterman, "Asymptotics and the Role of Minimal Models," *British Journal for the Philosophy of Science*, LIII (2002) 21–38 See also Batterman, *The Devil in the Details* (New York Oxford, 2001)

[13] Hartmann, "Idealization in Quantum Field Theory," in Niall Shanks, ed , *Idealization in Contemporary Physics* (Atlanta, GA Rodopi, 1998), pp 99–122

[14] Cartwright distinguishes this view from what she calls idealization, which is closer to Galilean idealization In a more recent defense of this distinction, Martin Jones has cogently argued that abstraction is best seen as a kind of omission, whereas idealization is the assertion of falsehood Cartwright's and Jones's proposal is perfectly reasonable—omission and distortion are distinguishable practices However, since I am arguing for pluralism about the nature of idealization, I see no reason why we should not treat minimalist modeling as a form of idealization See Jones, "Idealization and Abstraction A Framework," in Jones and Cartwright, eds , *Idealization XII Correcting the Model Idealization and Abstraction in the Sciences* (New York Rodopi, 2005), pp 173–217, for a careful defense of the alternative view Also see Humphreys, "Abstract and Concrete," *Philosophy and Phenomenological Research*, LV (1995) 157–61, for a criticism of Cartwright's view and an argument that idealization (in Cartwright's sense) will almost always come along with abstraction in real scientific contexts

[15] Cartwright, *Nature's Capacities*, p 187

of the vibrational properties of a covalent bond, but one that is commonly used in spectroscopy Galilean idealizers would justify the use of this model by saying that it is pragmatically useful for calculating energies, thus avoiding having to calculate the many-dimensional potential energy surface for the whole molecule Minimalist idealizers, however, would justify the use of this model by suggesting that it captures what really matters about the vibrations of covalent bonds The extra detail in the full potential energy surface, they would argue, is extraneous

As this example illustrates, the most important differences between Galilean and minimalist idealization are the ways that they are justified Even when they produce the same representations, they can be distinguished by the rationales they give for idealization Further, while Galilean idealization ought to abate as science progresses, this is not the case for minimalist idealization Progress in science and increases in computational power should drive the two apart, even if they generate the same model at a particular time

Just as there is no single account of minimalist idealization, there is no single account of its justification However, all of the influential accounts described above agree that minimalist idealization should be justified with respect to the cognitive role of minimal models they aid in scientific explanations Hartmann argues that minimal models literally tell us how phenomena behave in a simpler world than our own This gives us the necessary information to explain real-world phenomena For Batterman, minimal models demonstrate how fundamental structural properties of a system generate common patterns among disparate phenomena Strevens and Cartwright look at things more causally, describing the role of minimal models as showing us the causal factors that bring about the phenomenon of interest In all of these cases, minimalist idealization is connected to scientific explanation Minimal models isolate the explanatorily causal factors either directly (Cartwright and Strevens), asymptotically (Batterman), or via counterfactual reasoning (Hartmann) In each case, the key to explanation is a special set of explanatorily privileged causal factors Minimalist idealization is what isolates these causes and thus plays a crucial role for explanation This means that unlike Galilean idealization, minimalist idealization is not at all pragmatic and we should not expect it to abate with the progress of science

Multiple-Models Idealization Multiple-models idealization (hereafter, MMI) is the practice of building multiple related but incompatible models, each of which makes distinct claims about the nature and causal structure giving rise to a phenomenon MMI is similar to minimalist idealization in that it is not justified by the possibility

of de-idealization back to the full representation However, it differs from both Galilean and minimalist idealization in not expecting a single best model to be generated This type of idealization is most closely associated with a distinctive kind of theorizing called *modeling*[16] or *model-based science*[17]

One most commonly encounters MMI in sciences dealing with highly complex phenomena In ecology, for example, one finds theorists constructing multiple models of phenomena such as predation, each of which contains different idealizing assumptions, approximations, and simplifications Chemists continue to rely on both the molecular orbital and valence bond models of chemical bonding, which make different, incompatible assumptions In a dramatic example of MMI, the United States National Weather Service employs three complex models of global circulation patterns to model the weather Each of these models contains different idealizing assumptions about the basic physical processes involved in weather formation Although attempts have been made to build a single model of global weather, the NWS has determined that the best way to make high fidelity predictions is to employ all three models, despite the considerable expense of doing so[18]

The literature about MMI is less well developed then the others, so there is less of a clear consensus about its justification But one especially important justification of MMI is the existence of *tradeoffs*, a position closely associated with biologist Richard Levins and his philosophical allies[19] This justification begins by noting that theorists have different goals for their representations, such as accuracy, precision, generality, and simplicity Levins further argues that these desiderata and others can trade off with one another in certain circumstances, meaning that no single model can have all of these properties to the highest magnitude If a theorist wants to achieve

[16] Weisberg, "Who Is a Modeler?"

[17] Peter Godfrey-Smith, "The Strategy of Model Based Science," *Biology and Philosophy*, XXI (2006) 725–40

[18] Details about the three primary models, as well as a number of others employed by the NWS can be found at http //www meted ucar edu/nwp/pcu2

[19] Levins, "The Strategy of Model Building in Population Biology," in Elliott Sober, ed , *Conceptual Issues in Evolutionary Biology* (Cambridge MIT, 1984), pp 18–27, Jay Odenbaugh, "Complex Systems, Trade-Offs and Mathematical Modeling A Response to Sober and Orzack," *Philosophy of Science*, LXX (2003) 1496–507, Weisberg, "Qualitative Theory and Chemical Explanation," *Philosophy of Science*, LXXI (2004) 1071–81, and "Forty Years of 'The Strategy' Levins on Model Building and Idealization," *Biology and Philosophy*, XXI (2006) 623–45, John Matthewson and Weisberg, "The Structure of Tradeoffs in Scientific Modeling" (manuscript) For a critique of these ideas, see Steven H Orzack and Sober, "A Critical Assessment of Levins' 'The Strategy of Model Building in Population Biology'," *Quarterly Review of Biology*, LXVIII (1993) 533–46

high degrees of generality, accuracy, precision, and simplicity, she will need to construct multiple models Levins summarizes his discussion of these issues as follows

> The multiplicity of models is imposed by the contradictory demands of a complex, heterogeneous nature and a mind that can only cope with few variables at a time, by the contradictory desiderata of generality, realism, and precision, by the need to understand and also to control, even by the opposing esthetic standards which emphasize the stark simplicity and power of a general theorem as against the richness and the diversity of living nature These conflicts are irreconcilable Therefore, the alternative approaches even of contending schools are part of a larger mixed strategy But the conflict is about method, not nature, for the individual models, while they are essential for understanding reality, should not be confused with that reality itself (*op cit*, p 431)

Our cognitive limitations, the complexity of the world, and constraints imposed by logic, mathematics, and the nature of representation, conspire against simultaneously achieving all of our scientific desiderata Thus, according to Levins, communities of scientists should construct multiple models, which collectively can satisfy our scientific needs

Several other justifications for MMI can be found in the literature Wimsatt argues that highly idealized models are important because, taken together, they help us develop truer theories [20] Population biologists Robert May and Joan Roughgarden argue that clusters of simple models increase the generality of a theoretical framework, which can lead to greater explanatory depth [21] Finally, Strevens's account of idealization can also be used to justify MMI For Strevens, a theorist first finds a minimal causal model for a phenomenon of interest She idealizes when she makes this highly abstract model more concrete, and in doing so introduces (nondifference making) distortions The processes of filling in the minimal causal model with concrete details can be carried out in different ways, hence this process can yield multiple, idealized models

Some of these motivations suggest strong parallels between MMI and minimalist idealization In some cases, one cannot build a single minimal model that contains all of the core causal factors for a class of phenomena Yet it may be possible, in such cases, to build a small set of models, each of which highlights a different factor and which

[20] Wimsatt, "False Models as a Means to Truer Theories," in M Nitecki and A Hoffmann, eds, *Neutral Models in Biology* (New York Oxford, 1987), pp 23–55

[21] Jonathan [Joan] Roughgarden, *Theory of Population Genetics and Evolutionary Ecology An Introduction* (New York Macmillan, 1979), May, *Stability and Complexity in Model Ecosystems* (Princeton University Press, 2001)

together account for all of the core causal factors This motivation for MMI is parallel to the motivation for minimalist idealization, even though the practice itself is different

However, there are additional motivations for engaging in MMI that do not parallel the motivation for minimalist idealization For example, modelers may engage in MMI strictly for the purpose of maximizing predictive power, as do the forecasters at the National Weather Service Another instance of MMI may involve building a set of models that gives maximum generality, at the expense of capturing all of the core causal factors Still another is the synthetic chemist or engineer's motivation for MMI to find the set of idealized models that is maximally useful for creating new structures There are thus many motivations for MMI Some are pragmatic, where scientists are focused on prediction and structure construction, while some are explanatory and nonpragmatic

MMI also gives a complex, mixed answer about the permanence of idealization as science progresses In some domains, MMI may abate with the progress of science The National Weather Service may one day discover a single model that makes optimal predictions However, if tradeoffs exist between theoretically important desiderata in a particular domain, then we should not expect MMI to abate with further progress These tradeoffs are consequences of logic and mathematics and thus present a permanent justification for MMI

From the discussion so far, it may seem that the literature on idealization describes a hodgepodge of disparate practices, leaving no hope for any further analysis of idealization simpliciter This worry is not without merit because the methods, goals, and justifications of these three forms of idealization are quite distinct Although a fully unified account of the three kinds of idealization is impossible, some progress can be made towards developing a unified framework with which to understand the practice of idealization in general This framework focuses on the *goals* associated with idealization, rather than the activities or products of it I call these goals the *representational ideals* of idealization

II REPRESENTATIONAL IDEALS

Representational ideals are the goals governing the construction, analysis, and evaluation of theoretical models They regulate which factors are to be included in models, set up the standards theorists use to evaluate their models, and guide the direction of theoretical inquiry Representational ideals can be thought of as having two components *inclusion rules* and *fidelity rules* Inclusion rules tell the theorist which kinds of properties of the phenomenon of interest, or *target*

system, must be included in the model, while fidelity rules concern the degrees of precision and accuracy with which each part of the model is to be judged

An important, albeit very simple, representational ideal is called COMPLETENESS, which is associated with classic accounts of scientific method As such, it forms an important background against which every kind of idealization can be discussed

COMPLETENESS According to COMPLETENESS, the best theoretical description of a phenomenon is a complete representation The relevant sense of 'completeness' has two components associated with its inclusion rules and fidelity rules, respectively The inclusion rules state that each property of the target phenomenon must be included in the model Additionally, anything external to the phenomenon that gives rise to its properties must also be included in the model Finally, structural and causal relationships within the target phenomenon must be reflected in the structure of the model COMPLETENESS's fidelity rules tell the theorist that the best model is one that represents every aspect of the target system and its exogenous causes with an arbitrarily high degree of precision and accuracy

The description of COMPLETENESS given so far is accurate, but potentially misleading With very few exceptions, the inclusion and fidelity rules of COMPLETENESS set a goal that is impossible to achieve Unless extremely self deceived, or in possession of an extremely simple and abstract target system, no theorist thinks that complete representation is actually possible Given the impossibility of achieving complete representation, how can COMPLETENESS play a guiding role in scientific inquiry?

Despite its unattainable demands, COMPLETENESS can guide inquiry in two ways First, COMPLETENESS sets up a scale with which one can evaluate all representations including suboptimal ones If a theorist wants to rank several representations of the same phenomenon and has adopted COMPLETENESS, she has a straightforward way to do so The closer a representation comes to completeness, the better it scores I call this the *evaluative function* of the representational ideal because it sets the standards for evaluating sub-optimal representations

The second and more important way that COMPLETENESS can guide inquiry is through its *regulative function* Regulative functions are similar to what Kant called *regulative ideals* They do not describe a cognitive achievement that is literally possible, rather, they describe a target or aim point They give the theorist guidance about what she should strive for and the proper direction for the advancement of her research program If a theorist adopts COMPLETENESS, she knows that she should always strive to add more detail, more complexity, and

more precision to her models This will bring her closer to the ideal of completeness, although she will never fully realize this goal

COMPLETENESS is a unique representational ideal because it directs theorists to include everything in their representations All other ideals will build in some aspect of approximation or distortion In thinking about ideals other than COMPLETENESS, we can begin to see the outline of a framework for characterizing the three kinds of idealization Different kinds of idealization will be associated with different representational ideals Before we carry this analysis forward, let us consider several additional representational ideals

SIMPLICITY After COMPLETENESS, the next most straightforward ideal is SIMPLICITY The inclusion rule for this ideal councils the theorist to include as *little* as possible, while still being consistent with the fidelity rules The fidelity rule for SIMPLICITY demands a qualitative match between the behavior of target system and the properties and dynamics of the model

SIMPLICITY is primarily employed by working scientists in two contexts [22] The first is pedagogical Students are often introduced to the simplest possible model that can make sense of the data, even where scientists believe that the model contains serious problems One example of this is in the Lewis electron pair model of chemical bonding This model is not even quantum mechanical, yet it can be used to account for many canonical molecular structures Beginning students are introduced to this model as a way of building intuitions about chemical structure and reactivity

The second scientific context where SIMPLICITY is employed is when theorists construct models to test general ideas "A minimal model for an idea tries to illuminate a hypothesis [It] is not intended to be tested literally, any more than one would test whether the models for a frictionless pulley or a frictionless inclined plane are wrong "[23] This second use represents a motivation and justification for a particular kind of modeling in scientific practice Theorists often begin a project by trying to determine what kind of minimal structures could generate a property of interest They do not need to know, at first,

[22] There is also a long tradition which investigates the epistemic role of simple models In some circumstances, it seems that simple models ought to be preferred because they are more likely to be true This is a different kind of justification for the use of simple models than I am discussing in this article For a recent defense of the possible epistemic significance of simplicity, see Malcom Forster and Sober, "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions," *British Journal for the Philosophy of Science,* XLV (1994) 1–35

[23] Joan Roughgarden, *Primer of Ecological Theory* (Upper Saddle River, NJ Prentice Hall, 1998), p x

how a specific target system actually works Once the dynamics are understood in simple models, theorists examine more complex models and empirical data to assess the plausibility of the simple model's explanation of a real system's behavior

1-CAUSAL This representational ideal instructs the theorist to include in the model only the core or primary causal factors that give rise to the phenomenon of interest Put in the language of the causation literature, this ideal tells the theorist to only include the factors that made a difference The theorist constructs a mathematical model of a much simpler system than the one actually being studied, one that excludes higher order causal factors These are the factors which make no difference to the occurrence of the phenomenon, but control the precise way in which the phenomenon occurs [24] This is closely related to SIMPLICITY, but unlike SIMPLICITY, 1-CAUSAL restricts the level of simplicity that is allowed If we are trying to construct the simplest possible model that can make predictions qualitatively compatible with our observations, there is no restriction on the kind or number of causal factors that must be included SIMPLICITY, for example, may allow us to neglect all quantum mechanical effects and use the Lewis model 1-CAUSAL, however, would not sanction the use of such a model because it requires the theorist to include the quantum mechanical interactions that compose the core physical explanation of the structure

1-CAUSAL's fidelity criteria make a considerable difference in determining when the theorist has constructed an adequate model because its inclusion rule (restriction to primary causal factors) is not very specific In addition, the fineness of specification of the target phenomenon itself will make a difference to the kind of model we can build Imagine that we wanted to build a 1-CAUSAL model for the maintenance of the sex ratio We would need a more complex model to explain the 1 05 1 ratio of male to female *Homo sapiens*, than if we only were interested in why the sex ratio is roughly 50 50 Even holding the fidelity criteria fixed, the best model would be different in these two cases, with the former requiring greater specification of internal and external causal factors

Models generated using 1-CAUSAL are especially useful in two contexts Like the models generated with SIMPLE, they can be used as starting points for the formulation and analysis of more complex models 1-CAUSAL models are typically generated when one has a

[24] Of course, which factors do and do not make a difference to the occurrence of a phenomenon must be judged with respect to how precisely the phenomenon is individuated

reasonably comprehensive understanding of how a system behaves, since knowing the primary causal factors that give rise to a phenomenon requires knowing quite a lot about the system Further modeling from this point is usually aimed at greater quantitative accuracy, not deeper fundamental understanding

The second context where 1-CAUSAL is especially important involves scientific explanation Several recent philosophical accounts of scientific explanation have pointed to the central role that primary causal factors—the factors that really make a difference—play in scientific explanation [25] Recent work on the cognitive psychology of explanation has also emphasized the crucial role that picking out central causal factors plays in people's judgments of explanatory goodness [26] In their methodological discussions, a number of other scientists have commented on this connection For example, chemist Roald Hoffmann emphasizes that " if understanding is sought, simpler models, not necessarily the best and predicting all observables in detail, will have value Such models may highlight important causes and channels "[27] These accounts all suggest that models generated with 1-CAUSAL seem to be at the heart of theorists' explanatory practices

MAXOUT We now move from an ideal which looks superficially like SIMPLICITY to one that looks superficially like COMPLETENESS, the ideal called MAXOUT This ideal says that the theorist should maximize the precision and accuracy of the model's output It says nothing, however, about how this is to be accomplished

One way to work towards this ideal is by constructing highly accurate models of every property and causal factor affecting the target This is the same approach taken in COMPLETENESS, although the goal of MAXOUT is to achieve maximum output precision and accuracy, not a complete representation A second option, one more commonly associated with MAXOUT, is to engage in *model selection,*[28] a process of using statistics to choose a functional form, parameter set, and parameter values which best fit a large data set The model selected by these techniques is then continually optimized as further data comes

[25] James Woodward, *Making Things Happen A Theory of Causal Explanation* (New York Oxford, 2003), Strevens, "The Causal and Unification Accounts of Explanation Unified—Causally "

[26] Tania Lombrozo, "The Structure and Function of Explanations," *Trends in Cognitive Science,* x (2006) 464–70

[27] Hoffmann, V I Minkin, and Barry K. Carpenter, "Ockham's Razor and Chemistry," *Bulletin de la Société Chimique de France,* CXXXIII (1996) 117–30

[28] Forster, "The New Science of Simplicity," in Arnold Zellner, Hugo Keuzenkamp, and Michael McAleer, eds , *Simplicity, Inference and Modelling* (New York Cambridge, 2001), pp 83–117

in Finally, MAXOUT also sanctions the use of black box models, the sort that have amazing predictive power, but for unknown reasons These may be discovered using model selection techniques, or may be discovered in a more serendipitous fashion

At first blush, it may seem unscientific to adopt an ideal that values predictive power over everything else Most scientists believe that their inquiry is aimed at more than raw predictive power While scientists want to know how a system will behave in the future, they also want an explanation of why it behaves the way that it does MAXOUT ensures that we will generate models which are useful for predicting future states of the target system, but gives no guarantee that the models will be useful for explaining the behavior of the system

Nevertheless, representations generated by MAXOUT have their place in scientific inquiry Explanation and prediction are clearly both important goals of scientists, but there is no reason that they must both be fulfilled with the same model Theorists can adopt a mixed representational strategy, using different kinds of models to achieve different scientific goals It may also be rational to elevate predictive power above all other considerations in some situations Following his reflection on the importance of simple models quoted above, Hoffmann argues that "If predictability is sought at all cost—and realities of marketplace and judgments of the future of humanity may demand this—then simplicity may be irrelevant "[29]

P-GENERAL Generality is a desideratum of most models This desideratum really has two distinct parts *a-generality* and *p-generality* A-generality is the number of actual targets a particular model applies to given the theorist's adopted fidelity criteria P-generality, however, is the number of possible, but not necessarily actual, targets a particular model captures [30] The representational ideal P-GENERAL says that considerations of *p-generality* should drive the construction and evaluation of theoretical models.

While a-generality may seem like the more important kind of generality, theorists are often interested in p-generality for several reasons P-general models can be part of the most widely applicable theoretical frameworks, allowing real and nonreal target systems to be compared P-generality is also often thought to be associated with explanatory power This can be seen in both the philosophical literature on explanation and in the comments of theorists An excellent example of

[29] Hoffmann, "Ockham's Razor and Chemistry "

[30] For further discussion, see Weisberg, "Qualitative Theory and Chemical Explanation," and Matthewson and Weisberg, "The Structure of Tradeoffs in Scientific Model Building "

this can be found in R.A. Fisher's discussion of modeling the nonactual He begins by quoting Arthur Eddington

> We need scarcely add that the contemplation in natural science of a wider domain than the actual leads to a far better understanding of the actual [31]

Fisher goes on to argue

> [for] a biologist, speaking of his own subject, [this] would suggest an extraordinarily wide outlook No practical biologist interest in sexual reproduction would be led to work out the detailed consequences experienced by organisms having three or more sexes, yet what else should he do if he wishes to understand why the sexes are, in fact, always two (*ibid*, pp vii–ix)?

The key to understanding this actual system, Fisher argues, is to understand a possible, but nonactual one In the behavior of this nonactual three-sex system lies the key to understanding why the two-sex system evolved Some recent philosophical accounts of scientific explanation also stress the importance of p-generality to explanation [32]

P-GENERAL can also play a subtler regulative role Instead of trying to understand specific targets, theorists may wish to understand fundamental relationships or interactions, abstracted away from real systems For example, ecologists may wish to study predation or competition, far removed from the interactions of particular species In such cases, P-GENERAL is often adopted, guiding theorists to develop models that can be applied to many real and possible targets This exploratory activity is a very important part of modern theoretical practice, although we do not yet have good philosophical account of how it works [33] One thing we do know, however, is that there is a delicate balance between achieving deep and insightful p-generality and low-fidelity, uninformative p-generality, generated by overly simplistic models

We have now looked at a number of representational ideals, the goals that guide theoretical inquiry As I mentioned at the beginning of this section, representational ideals are at the core of the practice

[31] Fisher, *The Genetical Theory of Natural Selection* (New York Oxford, 1930), pp viii–ix, quoting Eddington's *The Nature of the Physical World.*

[32] Strevens, "The Causal and Unification Accounts", Woodward, *Making Things Happen*

[33] Some aspects of this exploratory mode of theorizing are discussed in Levins, "The Strategy of Model Building", Wimsatt, "Robustness, Reliability and Overdetermination," in M Brewer and B Collins, eds, *Scientific Inquiry and the Social Sciences* (San Francisco Jossey-Bass, 1981), pp 124–63, Weisberg, "Robustness Analysis," *Philosophy of Science,* LXXIII (2006) 730–42, Patrick Forber, "On Biological Possibility and Confirmation" (unpublished manuscript)

of idealization and a systematic account of them can ultimately lead us to a more unified understanding of idealization To that end, we now turn back to the three kinds of idealization and consider which representational ideals are associated with them

III IDEALIZATION AND REPRESENTATIONAL IDEALS

Recall that Galilean idealization is the practice of introducing distortions into theories in order to simplify them and make them computationally tractable It is justified pragmatically, introduced to make a model more computationally tractable, but with the ultimate intention of de-idealizing, removing any distortion, and adding detail back to the model Models generated by Galilean idealization are thus approximate, but carry with them the intention of further revision, ultimately reaching for a more precise, accurate, and complete model The ultimate goal of Galilean idealization is complete representation, its representational ideal is thus COMPLETENESS

Minimalist idealizers are not interested in generating the most truthful or accurate model Rather, they are concerned with finding minimal models, discovering the core factors responsible for the target phenomenon Minimalist idealizers thus adopt the representational ideal 1-CAUSAL, the ideal that says the best model is the one that includes the primary causal factors that account for the phenomenon of interest, up to a suitable level of fidelity chosen by the theorist While minimalist idealizers may sometime look like they are adopting SIMPLICITY, this is almost always inaccurate, because theorists engage in minimalist idealization to really understand how the target phenomena work and why they behave the way that they do This requires finding the causal factors that really do make a difference, not a model that simply can reproduce the phenomenon qualitatively

Like Galilean idealization's representational ideal, minimalist idealization's ideal also demands the construction of a single model for a particular target or class of target phenomena One typically engages in minimalist idealization in order to generate explanatory models Such models tend to be ones that simultaneously unify many target phenomena into a class and identify the causal factors which really make a difference For the class of phenomenon of interest, this will mean finding a single model, despite the fact that it will leave out quite a lot of detail which accounts for the uniqueness of each target

Finally, we can consider MMI The biggest difference between MMI and the other kinds of idealization is that there is no single representational ideal which is characteristic of it Pretty much any representational ideal—including 1-CAUSAL and in rare cases

COMPLETENESS—can play a role in this form of idealization MMI arises because of the existence of tradeoffs between different theoretical desiderata This suggests that not all desiderata are simultaneously maximizable, at least in a single model Thus the most significant aspect of MMI is that it instructs theorists to construct a series of models which pursue different desiderata and are guided by multiple representational ideals

Consider, for example, the ecological research program that is concerned with understanding predation A cursory look at the ecological literature on predation, reveals little in the way of the search for a single, best model of predation Instead, one finds a series of models, some of which are more precise and accurate, some of which are more qualitative, some of which are very well suited for populations that are homogenously distributed in space, and some of which are flexible enough to deal with complex spatial structure This situation is the norm in theoretical ecology As John Maynard Smith explained, "For the discovery of general ideas in ecology different kinds of mathematical description, which may be called models, are called for "[34]

For modern ecologists pursuing MMI, a full understanding of the ecological world is going to depend on multiple, overlapping, possibly incompatible models How might we justify this kind of pluralism? One possible approach is anti-realist We could argue that maximizing empirical adequacy in some cases requires the use of multiple models Since anti-realism only requires that models be empirically adequate, the use of different kinds of idealized models is unproblematic

This line of response is available to anti-realists, but neglects some of the motivations for building multiple models that theorists have discussed in the literature The same ecologists who champion the use of multiple models very explicitly describe this practice as aimed at having a more complete understanding of the phenomena of interest, not simply making accurate predictions As Levins puts it, "[O]ur truth is at the intersection of independent lies" (*op cit*, p 20) This is clearly a realist sentiment To understand if it is justified we must ask whether the use of multiple idealized models, or the use of any idealized models at all, is compatible with scientific realism

IV IDEALIZATION, REPRESENTATIONAL IDEALS, AND THE AIMS OF SCIENCE

Peter Godfrey-Smith gives the following helpful formulation of scientific realism "One actual and reasonable aim of science is to give us accurate descriptions (and other representations) of what reality

[34] Maynard Smith, *Models in Ecology* (New York Cambridge, 1974), p 1

is like This project includes giving us accurate representations of aspects of reality that are unobservable "[35] The realist thus believes that scientists aim and sometimes succeed at representing this external, independent reality, while anti-realists demur, at least when it comes to unobservables

Prima facie, idealization looks like it might cause problems for scientific realism All three forms of idealization involve the willful distortion of scientific representations Willful distortion and approximation appears to militate against Godfrey-Smith's conception of realism, because the theorist is not even aiming to give an accurate description of what mind-independent reality is like Despite this, I think all three kinds of idealization are compatible with the sort of realism sketched by Godfrey-Smith, if his definition is understood in a broad and sophisticated way

Galilean idealization is the most straightforwardly compatible with scientific realism Galilean idealizers often fall short of their representational ideal of COMPLETENESS and may even do so willingly However, in the long run, the Galilean idealizer does aim to give complete, nondistorted, perfectly accurate representations In order to accommodate the possibility of Galilean idealization, scientific realists need to understand that achieving accurate representations of complex phenomena is an ongoing process Even when the short-term practice involves the willful introduction of distortion, the long-term aim can still be to give an accurate representation of what reality is really like Thus scientific realism is perfectly compatible with Galilean idealization, if the realist aim is understood to be long-term or ultimate

Minimalist idealization and MMI present more serious challenges to scientific realism It will not be possible for minimalist and MMI idealizers to assent to at least one interpretation of Godfrey-Smith's formulation because they do not ever aim to give a *fully* accurate representation of reality However, defenders of minimalist idealization aim to uncover real causal structure, or fundamental patterns in common between multiple phenomena This suggests that a weaker reading of Godfrey-Smith's formulation, which does not require fully accurate representations, is compatible with minimalist idealization

There are other respects in which minimalist idealization is compatible, and indeed demands a kind of realism Consider the goals and justification of minimalist idealization Minimalist idealizers are trying to model the most important causal factors that underlie the

[35] Godfrey-Smith, *Theory and Reality* (Chicago University Press, 2003), p 176

properties and behaviors of target phenomena That is, they often recognize that real scientific explanation involves the identification of the core causal factors giving rise to the system, not all of the details This recognition is surely a realist one While minimalist idealizers are decidedly not interested in the truth, the whole truth, and nothing but the truth, they want to know the truth about what really matters For their explanatory interests, representation of just a few key factors is what matters This representation must be accurate

Finally, consider multiple models idealization As this constitutes a more diverse set of practices, it is much harder to make a unified judgment about the degree of realism embodied by MMI Some kinds of representational ideals are clearly not realist For example, the ideal MAXOUT tells the theorist that she should seek maximal precision and accuracy in the output of her model However, this ideal provides no guidance about the internal structure of the model and is compatible with black-box models MAXOUT is also compatible with models that are willfully distorted with the sole aim of making the predictions more accurate So clearly a practice of idealization that only uses MAXOUT is incompatible with realism In such a case, theorists do not aim to give accurate representations of the underlying reality of their target phenomena

As I described it, however, MMI transcends relying on any one kind of representational ideal It is a strategy for investigating phenomena when complexity and tradeoffs preclude the accomplishment of this in a single model When a theorist chooses to engage in MMI because her system is complex, but nevertheless wants to develop an accurate description of her target phenomenon, she is acting in a realist fashion At least one aim of her practice is the development of a mind-independent picture of a real-world phenomenon This attitude is also clearly realist in spirit, despite the fact that it will be strictly incompatible with Godfrey-Smith's definition of realism

Indeed, there can be several realist dimensions of MMI Typical episodes of MMI employ a package of representational ideals, which taken together, aim to give an accurate representation of real world systems While no single model may contain the complete picture of the properties and behavior of a complex system, a collection of them can Levins and Maynard Smith advocate the use of multiple models precisely because a collection of ecological models will give a more accurate representation of the behavior of real world ecosystems then any single one does

The recognition that multiple models can give a more accurate and informative representation of real world systems is itself another realist dimension of MMI, in fact one might call it a *higher order* realist

motivation for MMI Multiple models idealization is justified by the existence of tradeoffs between theoretically important desiderata such a simplicity, accuracy, precision, and generality If these tradeoffs exist in the way many scientists and philosophers believe that they do, adopting theoretical strategies that recognize them is the proper realist response Where tradeoffs exist, the realist should not be content to choose a single, most accurate model Such a strategy ignores important discoveries about the world, in this case about our representational capacities When faced with tradeoffs and complex systems, the realist should surely follow the Levinsonian strategy of multiple-model use

The goal of this discussion of realism is to show that, despite *prima facie* concerns about the incompatibility of idealization and realism, all three kinds of idealization can be made compatible with sophisticated forms of realism A more detailed study of these issues would look more carefully at each representational ideal and consider the extent to which its fidelity criteria and inclusion rules are compatible with realism and other scientific desiderata

What, then, have we learned about idealization and its justification? I have endeavored to show that the three kinds of idealization recognized in the philosophical literature are not competitors, but reflect three practices important to scientific inquiry What distinguishes them is not the product of their application, but rather the representational ideals which guide theorists in using them There is no single, over-arching justification for idealization Differing representational ideals respond to the demands of a complex world in different ways This precludes a single justification for idealization

MICHAEL WEISBERG

University of Pennsylvania

INDEX TO VOLUME CIV

INDEX OF AUTHORS
Including Contributing Authors and Reviewers

Aizawa, Kenneth, Understanding the Embodiment of Perception **104** (1) 5–25 '07

Barnes, Jonathan, Review of Bernard Williams *The Sense of the Past Essays in the History of Philosophy* **104** (10) 540–45 '07

Burge, Tyler, Predication and Truth **104** (11) 580–608 '07

Campbell, Richmond, What Is Moral Judgment? **104** (7) 321–49 '07

Clark, Andy, Curing Cognitive Hiccups A Defense of the Extended Mind **104** (4) 163–92 '07

Danks, David, Reasons as Causes in Bayesian Epistemology By Clark Glymour and David Danks **104** (9) 464–74 '07

Gillett, Carl, Understanding the New Reductionism The Metaphysics of Science and Compositional Reduction **104** (4) 193–216 '07

Glymour, Clark, Reasons as Causes in Bayesian Epistemology By Clark Glymour and David Danks **104** (9) 464–74 '07

Godfrey-Smith, Peter, Conditions for Evolution by Natural Selection **104** (10) 489–516 '07

Gold, Natalie, Collective Intentions and Team Agency By Natalie Gold and Robert Sugden **104** (3) 109–37 '07

Hansson, Sven Ove, The False Dichotomy between Coherentism and Foundationalism **104** (6) 290–300 '07

Hare, Caspar, Self-Bias, Time-Bias, and the Metaphysics of Self and Time **104** (7) 350–73 '07

Klein, Colin, An Imperative Theory of Pain **104** (10) 517–32 '07

Lear, Jonathan, Review of Bernard Williams *Philosophy as a Humanistic Discipline* **104** (10) 546–50 '07

Levin, Michael, Compatibilism and Special Relativity **104** (9) 433–63 '07

Manley, David, Safety, Content, Apriority, Self-Knowledge **104** (8) 403–23 '07

McPherson, Lionel K., Normativity and the Rejection of Rationalism **104** (2) 55–70 '07

Mendola, Joseph, Knowledge and Evidence **104** (3) 157–60 '07

Moller, Dan, Love and Death **104** (6) 301–16 '07

Moya, Carlos J, Moral Responsibility without Alternative Possibilities? **104** (9) 475–86 '07

Mumford, Stephen, Review of Peter Unger *All the Power in the World* **104** (8) 424–31 '07

Needham, Paul, Macroscopic Mixtures **104** (1) 26–52 '07

Okasha, Samir, Rational Choice, Risk Aversion, and Evolution **104** (5) 217–35 '07

Petersson, Bjorn, Collectivity and Circularity **104** (3) 138–56 '07

Pippin, Robert B, Review of Bernard Williams *In the Beginning Was the Deed Realism and Moralism in the Political Argument* **104** (10) 533–39 '07

Price, Huw, Naturalism **104** (8) 375–402 '07

Pust, Joel, Cartesian Knowledge and Confirmation **104** (6) 269–89 '07

Rattan, Gurpreet, The Two Worlds of Deflationism **104** (12) 609–38 '07

Reed, Baron, The Long Road to Skepticism **104** (5) 236–62 '07

Rothschild, Daniel, Presuppositions and Scope **104** (2) 71–106 '07

Sugden, Robert, Collective Intentions and Team Agency By Natalie Gold and Robert Sugden **104** (3) 109–37 '07

Vasiliou, Iakovos, Review of Gabriel Richardson Lear *Happy Lives and the Highest Good An Essay on Aristotle's Nicomachean Ethics* **104** (5) 263–68 '07

Waters, C Kenneth, Causes That Make a Difference **104** (11) 551–79 '07

Weisberg, Michael, Three Kinds of Idealization **104** (12) 639–59 '07

INDEX OF BOOKS REVIEWED
Listed by Author of Item Reviewed

Davidson, Donald *Truth and Predication*—Tyler Burge **104** (11) 580–608 '07

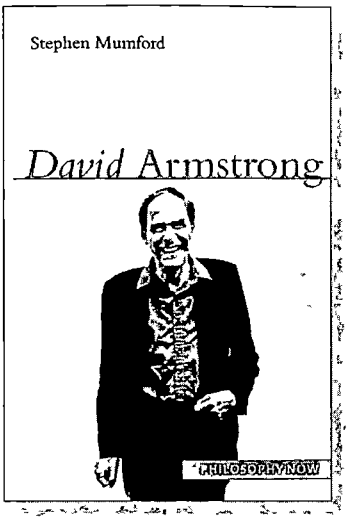
Lear, Gabriel Richardson *Happy Lives and the Highest Good An Essay on Aristotle's Nicomachean Ethics*—Iakovos Vasiliou **104** (5) 263–68 '07

Unger, Peter *All the Power in the World*—Stephen Mumford **104** (8) 424–31 '07

Williams, Bernard *In the Beginning Was the Deed Realism and Moralism in the Political Argument*—Robert B Pippin **104** (10) 533–39 '07

Williams, Bernard *Philosophy as a Humanistic Discipline*—Jonathan Lear **104** (10) 546–50 '07

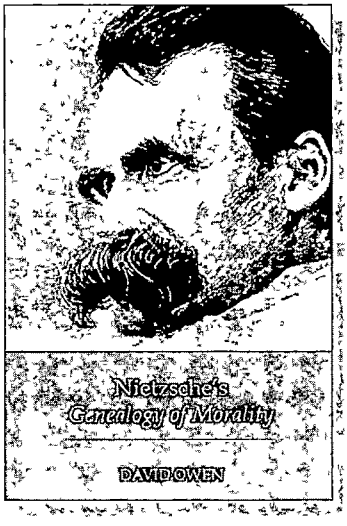
Williams, Bernard *The Sense of the Past Essays in the History of Philosophy*—Jonathan Barnes **104** (10) 540–45 '07



David Armstrong
Stephen Mumford

978-0-7735 3331-8 $22 95 paper
978-0-7735-3330 1 $80 00 cloth

"A well-organised, lucid, and accurate account of all the main aspects of Armstrong's philosophy A very good book "
– Keith Campbell, University of Sydney

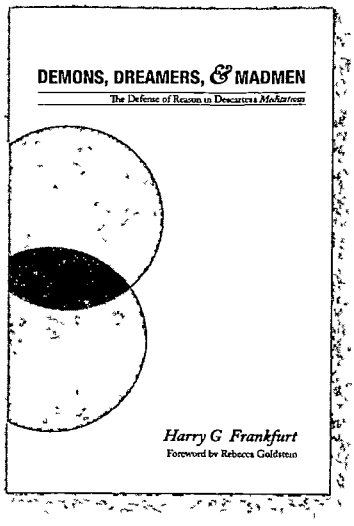


Nietzsche's Genealogy of Morality
David Owen

978-0-7735-3349-3 $22 95 paper
978 0-7735-3348-6 $80 00 cloth

"David Owen has established himself in recent years as one of the most insightful readers of Nietzsche In this book he not only sheds new light on Nietzsche's understandings of morality, but he shows conclusively Nietzsche's debt and differences to other moral and ethical theories "
– Tracy B Strong, University of California, San Diego

MQ UP
McGILL-QUEEN'S UNIVERSITY PRESS
www mqup ca


DEMONS, DREAMERS, & MADMEN
The Defense of Reason in Descartes's Meditations
Harry G Frankfurt
Foreword by Rebecca Goldstein

With a new foreword by Rebecca Goldstein and a new preface by the author

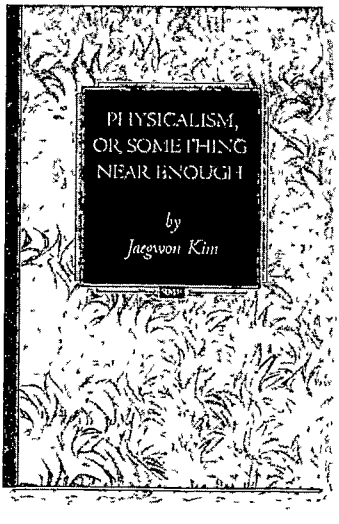
Demons, Dreamers, and Madmen

The Defense of Reason in Descartes's *Meditations*

Harry G Frankfurt

"One of the smartest books on Descartes in the literature in any language But it is not only an important contribution to the history of philosophy; Frankfurt also gives us new perspectives on knowledge and truth that have remained as fresh and lively as when they were first published"
—Dan Garber,
Princeton University

Paper $19 95 978 0 691-13416-1
Cloth $55 00 978-0-691-13415-4
Due January



New in paperback

Physicalism, or Something Near Enough

Jaegwon Kim

"This is an excellent book by one of the world's best philosophers working on the metaphysics of mind Nobody who is interested in the mind-body problem should miss the opportunity to study this book carefully"
—Jesper Kallestrup,
Philosophical Quarterly

Princeton Monographs in Philosophy
Harry G Frankfurt, Series Editor

Paper $16 95 978 0-691-13385-0
Due February

PRINCETON UNIVERSITY PRESS

New from The MIT Press

Folk Psychological Narratives

The Sociocultural Basis of Understanding Reasons

Daniel D Hutto

"This is a fresh, timely, and thought-provoking book Daniel Hutto has succeeded admirably in laying out a comprehensive alternative to 'theory of mind' approaches to psychological understanding"
— Peter Hobson, University of London

352 pp , $38 cloth

Time and Realism

Metaphysical and Antimetaphysical Perspectives

Yuval Dolev

"The great achievement of *Time and Realism* is that it shows how the two conflicting ideas of time misrepresent the role that temporality plays in our lives as well as in science This is contemporary philosophy at its best" — Hilary Putnam, Harvard University

264 pp , 3 illus , $30 paper

The Musical Representation

Meaning, Ontology, and Emotion

Charles O Nussbaum

"An extraordinarily rich book, which uses the tools of analytic philosophy and the resources of cognitive science to explore the delightful mystery that is human musical experience" — Lawrence Zbikowski, University of Chicago, author of *Conceptualizing Music*

392 pp , 50 illus , $38 cloth

Things and Places

How the Mind Connects with the World

Zenon W Pylyshyn

"This is one of the most provocative and challenging books to have appeared in philosophy or psychology in the last twenty years" — John Campbell, University of California, Berkeley

280 pp , 26 illus , $34 cloth

Integrating Evolution and Development

From Theory to Practice

edited by Roger Sansom and Robert Brandon

Scholars argue for the importance of the developmental synthesis, or evo-devo, discussing the history and potential of this growing field of study and presenting specific case studies

336 pp , 45 illus , $34 paper

The Letters of George Santayana

Book Eight, 1948-1952

edited and with an introduction by William G Holzberger

Letters from the last years of Santayana's life, written as he completed *Dominations and Powers*, the final volume of his autobiography, and the one-volume abridgement of his early five-part masterwork, *The Life of Reason*

704 pp , $70 cloth

Evolutionary Psychology as Maladapted Psychology

Robert C Richardson

A philosopher subjects the claims of evolutionary psychology to the evidential and methodological requirements of evolutionary biology, concluding that evolutionary psychology's explanations amount to speculation disguised as results

248 pp , 13 illus , $30 cloth

Situating Semantics

Essays on the Philosophy of John Perry

edited by Michael O'Rourke and Corey Washington

Original commentary on the work of philosopher John Perry by prominent contemporary analytic philosophers, with Perry's detailed and original responses

608 pp , 6 illus , $40 paper

The MIT Press To order call 800-405-1619 • http //mitpress mit edu

Moral Psychology

Volume 1 The Evolution of Morality Adaptations and Innateness

Volume 2 The Cognitive Science of Morality Intuition and Diversity

Volume 3 The Neuroscience of Morality Emotion, Brain Disorders, and Development

edited by Walter Sinnott-Armstrong

"Represents a powerful and penetrating discussion of the emerging science of moral behavior With these volumes, Sinnot-Armstrong is revealed as a leading mind in this field" — Michael Gazzaniga, University of California, Santa Barbara

Volume 1 • 496 pp , 7 illus , $30 paper
Volume 2 512 pp , 6 illus , $30 paper
Volume 3 504 pp , $30 paper

Distributed Cognition and the Will

Individual Volition and Social Context

edited by Don Ross, David Spurrett, Harold Kincaid, and G Lynn Stephens

Philosophers and behavioral scientists discuss what, if anything, of the traditional concept of individual conscious will can survive recent scientific discoveries that human decision-making is distributed across different brain processes and through the social environment

368 pp , 22 illus , $34 paper

Ontology of Consciousness

Percipient Action

edited by Helmut Wautischer
foreword by Robert A F Thurman

"Scholars who study consciousness, whether they be behavioral, social or biological scientists, or just educated readers, will find in this volume a store of data necessary for the pursuit of this subject" — Douglass Price-Williams, University of California, Los Angeles

672 pp , 73 illus , $38 paper

ZONE Books

Objectivity

Lorraine Daston and Peter Galison

The emergence of objectivity in the mid-nineteenth-century sciences, as revealed through images in scientific atlases—a story of how lofty epistemic ideals fuse with workaday practices

Distributed for Zone Books 500 pp , 108 black & white illus , 32 color illus , $38 95 cloth

Vital Nourishment

Departing from Happiness

François Jullien
translated by Arthur Goldhammer

A philosophical inquiry into how to "feed life," or nourish it, draws from early Chinese thinker Zhuanghi to explore notions of breath, energy, and immanence

Distributed for Zone Books 170 pp , $25 95 cloth

Profanations

Giorgio Agamben
translated by Jeff Fort

Pivotal essays by a provocative Italian philosopher on memory and oblivion, on what is lost and what remains

Distributed for Zone Books 200 pp , 6 illus , $25 95 cloth

now in paperback

In Praise of Blandness

Proceeding from Chinese Thought and Aesthetics

François Jullien
translated by Paula M Varsano

A consideration of blandness not as the absence of defining qualities but as the harmonious union of all potential values—an infinite opening into human experience

Distributed for Zone Books 152 pp., 8 illus., $18 95 paper

now in paperback

Things that Talk

Object Lessons from Art and Science

edited by Lorraine Daston

"This collection is a feast for students of art, modern Western history, and philosophy" — *Library Journal*

Distributed for Zone Books 250 pp , 9 color illus , 82 black & white illus , $21 95 paper

New books from YALE

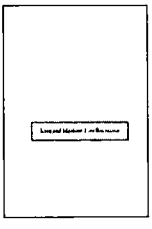
Leo Strauss
An Intellectual Biography
DANIEL TANGUAY

"Tanguay's intellectual biography of Leo Strauss is **the fullest and fairest analysis** of Leo Strauss's thought available in any language "
—Daniel J Mahoney
$30 00

Existentialism Is a Humanism
JEAN-PAUL SARTRE
Introduction by Annie Cohen-Solal
Preface and Notes by Arlette Elkaim-Sartre

This book presents new English translations of Sartre's "Existentialism Is a Humanism" and his analysis of Camus's *The Stranger*, along with a discussion of these works by acclaimed Sartre biographer Annie Cohen-Solal
$9 95 **paperback**

Kant and Idealism
TOM ROCKMORE



"This much-needed book covers in detail one of the great 'black holes' of contemporary philosophical discussion "
—Roberto Poli
$55 00

What Is Sport?
ROLAND BARTHES
Translated by Richard Howard

"Full of paradoxes, surprising rapprochements, and the melancholy wisdom of which Barthes was always a master "—Peter Starr
$15 00 **paperback**

The Persistence of Poverty
Why the Economics of the Well-Off Can't Help the Poor
CHARLES KARELIS

"This **astonishingly lively** [book] will make economists, philosophers, and ordinary engaged citizens rethink just about everything they took for granted about the causes and cures of poverty "—Alan Ryan
$30 00

The Leibniz-Des Bosses Correspondence
Translated, Edited, and with an Introduction by BRANDON C LOOK and DONALD RUTHERFORD

"A detailed and historically informed picture of the Leibniz-Des Bosses correspondence as well as a brilliant introductory essay that will be hugely influential on Leibniz scholarship "—Richard T W Arthur
The Yale Leibniz Series
$100 00

YALE University Press • yalebooks.com

New books from YALE

Thinking Politically

Essays in Political Theory
MICHAEL WALZER
Selected, Edited, and with an Introduction by David Miller

This collection **brings together some of Walzer's most important work** and includes a substantial introduction by David Miller that connects his ideas to wider currents of political thought
$30 00

Babies by Design

The Ethics of Genetic Choice
RONALD M GREEN

"Green presents ethical positions and arguments in the context of modern thought, and in a language that will be accessible to scholars, scientists and lay-people "—Lee Silver
$26 00

Thinking in Circles

An Essay on Ring Composition
MARY DOUGLAS

"Douglas's mission is to rescue great works of art from the trash heap in which they may have been thrown for lack of understanding, precisely, of their ring structure **In this she certainly succeeds."**—Leonard C Muellner
$28 00

New in Paper

Manliness

HARVEY C MANSFIELD

"A deeply creative work, one which endeavors to reinterpret Western political philosophy in terms of its chosen theme—a bold as well as an innovative project Important and thoughtful "
—James Piereson, *New Criterion*
$17 00 **paperback**

A Guide to Philosophy in Six Hours and Fifteen Minutes

WITOLD GOMBROWICZ
Translated by Benjamin Ivry

"Gombrowicz gives philosophy **witty and insightful** turns in this nutshell form *The Guide* provid[es] readers prompts to philosophize on their own, taking up and developing this necessary activity for themselves "
—James P Hammersmith, *Southern Humanities Review*
$15 00 **paperback**

Second Nature

Brain Science and Human Knowledge
GERALD M EDELMAN, M D , PH D

"A remarkable contribution to the philosophy of the mind [that] breaks new ground to an age-old problem by launching brain-based epistemology Original, lucid, concise, succinct easily the best in the field "—Apostolos P Georgopoulos
$13 00 **paperback**

YALE University Press • yalebooks com

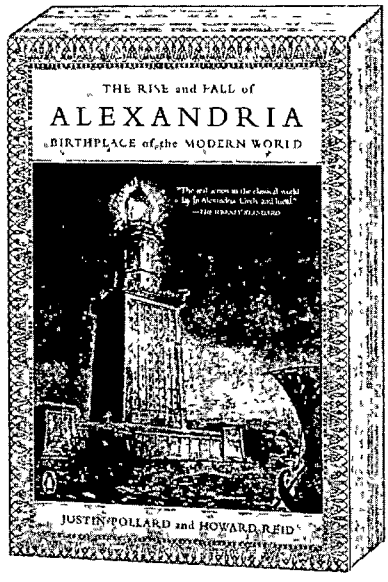
JUSTIN POLLARD & HOWARD REID

THE RISE AND FALL OF

ALEXANDRIA

BIRTHPLACE of the MODERN WORLD

A short history of nearly everything classical



"Tell[s] the story of ancient Alexandrian science with relish and brio"
— *The Weekly Standard*

"A marvelous work, a sweeping narrative, as grand as it is unusual, Pollard and Reid breathe life into the dust of philosophers and kings"
— Barry Strauss, author of *The Battle of Salamis*

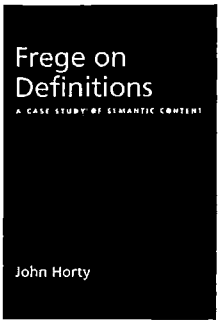
"Lucid, entertaining, full of exquisite images"
— Michael Hirst, screenwriter of *Elizabeth The Golden Age*

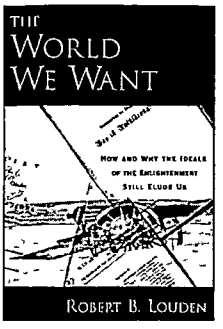
In this short narrative history, Justin Pollard and Howard Reid bring one of history's most fascinating and prolific cities to life, creating a treasure trove of our intellectual and cultural origins Famous for its lighthouse, its library, and its fertile intellectual and spiritual life—it was here that Christianity came to prominence as a world religion—Alexandria now takes its rightful place alongside Greece and Rome as a foundation of the modern world

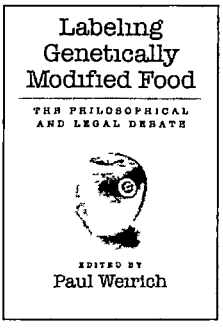
Penguin 352 pp 978-0-14-311251-8 $16.00

www.penguin.com/academic PENGUIN GROUP (USA)
Academic Marketing Department 375 Hudson St. New York, NY 10014

New From OXFORD







FREGE ON DEFINITIONS
A Case Study of Semantic Content
JOHN HORTY

In this short monograph, John Horty explores the difficulties presented for Gottlob Frege's semantic theory, as well as its modern descendents, by the treatment of defined expressions
2007 176 pp $55 00

THE WORLD WE WANT
How and Why the Ideals of the Enlightenment Still Elude Us
ROBERT B LOUDEN

The World We Want compares the future world that Enlightenment intellectuals had hoped for with our world at present Louden advances an argument that the widespread Enlightenment assumption that external institutional changes lead to internal attitudinal changes is not well supported by historical experience—and that many of the ideals advocated by the Enlightenment do not fit with human nature
2007 344 pp $65 00

PHYSICAL REALIZATION
SYDNEY SHOEMAKER

How can physicalism be true? Sydney Shoemaker's answer to this question involves showing how the mental properties of a person can be "realized in" the physical properties of that person, and how all instantiations of properties in macroscopic objects can be realized in microphysical states of affairs *Physical Realization* will be essential reading for many philosophers in the fields of metaphysics and the philosophy of mind
2007 160 pp $34 00

ACROSS THE BOUNDARIES
Extrapolation in Biology and Social Science
DANIEL STEEL

In *Across the Boundaries*, Steel argues that previous accounts of extrapolation are inadequate and proposes a better approach that is able to answer methodological critiques of extrapolation from animal models to humans
(Environmental Ethics and Science Policy Series)
2007 256 pp $65 00

PHILOSOPHERS ON MUSIC
Experience, Meaning, and Work
Edited by KATHLEEN STOCK

Philosophers on Music Experience, Meaning, and Work presents significant new contributions to central issues in the philosophy of music, written by leading philosophers working in the analytic tradition Aestheticians, musicologists, music practitioners, and those interested in philosophy generally will find the papers in this volume rewarding reading
2007 236 pp , 25 musical examples $60 00

LABELING GENETICALLY MODIFIED FOOD
The Philosophical and Legal Debate
Edited by PAUL WEIRICH

Many countries require that food labels announce genetically modified ingredients The United States does not require such labeling Which labeling policy is best? This interdisciplinary book surveys various labeling policies and the cases for them It is the first comprehensive treatment of the debate about labeling genetically modified food
2007 272 pp $65 00

OXFORD UNIVERSITY PRESS

Prices are subject to change and apply only in the US To order please contact customer service at 1-866-445-8685, or visit us online at www oup com/us

CAMBRIDGE

New and Noteworthy Titles

The Culture of Vengeance and the Fate of American Justice
Terry Kenneth Aladjem
$85 00 Hb 978 0 521-88624 6 264 pp
$26 99 Pb 978-0 521-71386-3

An Introduction to Many-Valued and Fuzzy Logic
Semantics, Algebras, and Derivation Systems
Merrie Bergmann
$90 00 Hb 978 0 521-88128 9. 344 pp
$36 99 Pb 978-0 521-70757 2

The Moral Demands of Memory
Jeffrey Blustein
$80 00 Hb 978 0 521-88330-6 392 pp
$24 99 Pb 978 0 521 70972-9

Carnap and Twentieth-Century Thought
Explication as Enlightenment
A W Carus
$95 00 Hb 978-0 521-86227-1 360 pp

The Cambridge Companion to Carnap
Edited by Michael Friedman and Richard Creath
Cambridge Companions to Philosophy
$85 00 Hb 978 0 521-84015 6 392 pp
$29 99 Pb 978 0 521-54945 5

Ethics and the Environment
An Introduction
Dale Jamieson
Cambridge Applied Ethics
$90 00 Hb 978 0-521-86421-3 234 pp
$29 99 Pb 978-0-521-68284 8

Aquinas, Aristotle, and the Promise of the Common Good
Mary M Keys
$26 99 Pb 978-0-521-72238 4 270 pp

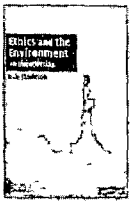
Second Edition!
The Economy of the Earth
Philosophy, Law, and the Environment
Mark Sagoff
$80 00 Hb 978 0 521 86755-9 280 pp
$27 99 Pb 978-0-521-68713 3

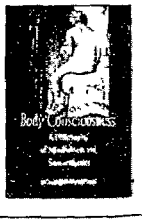
Body Consciousness
A Philosophy of Mindfulness and Somaesthetics
Richard Shusterman
$85 00 Hb 978-0 521-85890-8 240 pp
$24 99 Pb 978-0-521 67587 1

I Was Wrong
The Meanings of Apologies
Nick Smith
$99 00 Hb 978 0-521 86552-4 304 pp
$24 99 Pb 978 0-521-68423-1

Medically Assisted Death
Robert Young
$90 00 Hb 978 0 521-88024-4 260 pp
$29 99 Pb 978-0-521-70616-2

Prices subject to change





www cambridge org/us/philosophy

CAMBRIDGE UNIVERSITY PRESS